On the Properties and Structure of Bordered Words and Generalizations

by

Daniel Gabric

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Computer Science

Waterloo, Ontario, Canada, 2022

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	Tero Harju Prof. emer., Dept. of Mathematics and Statistics, University of Turku
Supervisor(s):	Jeffrey O. Shallit Professor, School of Computer Science, University of Waterloo
Internal Member:	Lila Kari Professor, School of Computer Science, University of Waterloo
Internal-External Member:	Jason Bell Professor, Dept. of Pure Mathematics, University of Waterloo
Other Member(s):	Daniel G. Brown Professor, School of Computer Science, University of Waterloo

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Combinatorics on words is a field of mathematics and theoretical computer science that is concerned with sequences of symbols called words, or strings. One class of words that are ubiquitous in combinatorics on words, and theoretical computer science more broadly, are the *bordered words*. The word w has a border u if u is a non-empty proper prefix and suffix of w. The word w is said to be bordered if it has a border. Otherwise w is said to be *unbordered*.

This thesis is primarily concerned with variations and generalizations of bordered and unbordered words.

In Chapter 1 we introduce the field of combinatorics on words and give a brief overview of the literature on borders relevant to this thesis.

In Chapter 2 we give necessary definitions, and we present a more in-depth literature review on results on borders relevant to this thesis.

In Chapter 3 we complete the characterization due to Harju and Nowotka of binary words with the maximum number of unbordered conjugates. We also show that for every number, up to this maximum, there exists a binary word with that number of unbordered conjugates.

In Chapter 4 we give results on pairs of words that almost commute and anti-commute. Two words x and y almost commute if xy and yx differ in exactly two places, and they anti-commute if xy and yx differ in all places. We characterize and count the number of pairs of words that almost and anti-commute. We also characterize and count variations of almost-commuting words. Finally we conclude with some asymptotic results related to the number of almost-commuting pairs of words.

In Chapter 5 we count the number of length-n bordered words with a unique border. We also show that the probability that a length-n word has a unique border tends to a constant.

In Chapter 6 we present results on factorizations of words related to borders, called *block palindromes*. A block palindrome is a factorization of a word into blocks that turns into a palindrome if each identical block is replaced by a distinct character. Each block is a border of a central block. We call the number of blocks in a block palindrome the *width* of the block palindrome. The *largest block palindrome* of a word is the block palindrome of the word with the maximum width. We count all length-n words that have a width-t *largest block palindrome*. We also show that the expected width of a largest block palindrome

tends to a constant. Finally we conclude with some results on another extremal variation of block palindromes, the *smallest block palindrome*.

In Chapter 7 we present the main results of the thesis. Roughly speaking, a word is said to be *closed* if it contains a non-empty proper border that occurs exactly twice in the word. A word is said to be *privileged* if it is of length ≤ 1 or if it contains a non-empty proper privileged border that occurs exactly twice in the word. We give new and improved bounds on the number of length-n closed and privileged words over a k-letter alphabet.

In Chapter 8 we work with a generalization of bordered words to pairs of words. The main result of this chapter is a characterization and enumeration result for this generalization of bordered words to multiple dimensions.

In Chapter 9 we conclude by summarizing the results of this thesis and presenting avenues for future research.

Acknowledgements

First and foremost, I would like to thank my supervisor Jeffrey Shallit, for his guidance, thoughtfulness, and patience throughout my time as a student. I would particularly like to thank him for inviting me to collaborate on interesting problems, while also encouraging me to pursue my own research interests as well.

I would also like to thank the members of my defence committee, Jason Bell, Daniel Brown, Tero Harju, and Lila Kari for their valuable time spent reviewing my thesis and attending my defence.

This thesis was supported in part by the David R. Cheriton graduate student scholarship and the Ontario graduate scholarship (OGS).

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Combinatorics on words is a field of mathematics and computer science that is concerned with sequences of symbols called words, or strings. This thesis is primarily concerned with variations of bordered [94] and unbordered words. A *word* (or *string*) is a sequence of symbols taken from a finite alphabet. The word w has a border u if $w = ux = yu$ for some words x, y. The word w is said to be *bordered* if it has a border u with $0 < |u| < n$. Otherwise w is said to be *unbordered*. Note that borders are allowed to overlap; see Example 1. Bordered and unbordered words are ubiquitous in combinatorics on words and appear in many computer science applications. For example, they occur in frame synchronization [75, 90, 50, 77], data compression [102], pattern matching [21, 52, 69], and more.

Example 1.
The French word `entente` is a bordered word. It has two borders `ente`, and `e`.
The English word `murmur` is also a bordered word. Its only border is `mur`.

Enumerating and proving properties about subsets of bordered and unbordered words has been of great interest in the field of combinatorics on words for a while. For example on the problem of enumeration, Nielsen's [80], and Lossers and Chapman's [20] recurrences for the number of unbordered and bordered words, Guibas and Odlyzko's [51] recurrences for the number of words that have a specific set of border lengths, Holub and Shallit's recurrence for the number of bordered words with a maximal border of size 1, and many more [50, 58, 84, 82].

Two additional well-known and intimately connected variations of bordered words are the closed words [37, 24] and the privileged words [67]. Roughly speaking, a word w is said

to be *closed* if it has a non-empty proper border u that occurs exactly twice in w. Similarly, a word is said to be *privileged* if it is of length ≤ 1 or if it contains a non-empty proper privileged border that occurs exactly twice in the word. A variation of closed words was first introduced by Gilbert [47] under the name *prefix-synchronized codes* in the field of frame synchronization; see Section 2.4 for the definition of *prefix-synchronized codes*. Guibas and Odlyzko [50] further studied these codes, enumerating them. Later Carpi and de Luca [25] independently discovered closed words, under the name *periodic-like words*. Since closed words and privileged words were defined in 2011 and 2013 a number of researchers have tried to enumerate the closed and privileged words. See [40, 79, 86, 87] for current bounds on the number of closed and privileged words.

Periodicity is another property that has been widely studied in the field of combinatorics on words. An integer p is said to be a *period* of the word $a_1a_2\cdots a_n$ if $a_i = a_{i+p}$ for all i, $1 \leq i \leq n-p$. It is known that a length-n word has a period p if and only if the word also has a border of length $n-p$. So periodicity and borderedness are very closely connected. For more on the connections between periodicity and borders see: the Ehrenfeucht-Silberger problem [36, 63, 64], Duval's conjecture [33, 34, 57, 61, 59, 35], Harju and Nowotka's results on border correlation [56, 60], and more.

In the next section we outline the main results and structure of this thesis.

1.1 Thesis outline

In Chapter 2 we give necessary definitions and we present a literature review on results on borders relevant to this thesis.

In Chapter 3 we complete the characterization due to Harju and Nowotka [56, 60] of binary words with the maximum possible number of unbordered conjugates. We also show that for any number, up to this maximum, there exists a binary word with that number of unbordered conjugates. We use the theorem-proving software `Walnut` written by Hamoon Mousavi [78] to aide us in the proofs. This work appears in [27].

In Chapter 4 we give results on pairs of words that almost commute and anti-commute. Two words x and y almost commute if xy and yx differ in exactly two places, and they anti-commute if xy and yx differ in all places. We characterize and count the number of pairs of words that almost and anti-commute. We also characterize and count variations of almost-commuting words. Finally we conclude with some asymptotic results related to the number of almost-commuting pairs of words. This work appears in [44].

In Chapter 5 we count the number of length-n bordered words with a unique border. We also show that the probability that a length-n word has a unique border tends to a constant.

In Chapter 6 we present results on factorizations of words related to borders, called *block palindromes*. A block palindrome is a factorization of a word into blocks that turns into a palindrome if each identical block is replaced by a distinct character. Each block is a border of a central block. We call the number of blocks in a block palindrome the *width* of the block palindrome. The *largest block palindrome* of a word is the block palindrome of the word with the maximum width. We count all length-n words that have a width-t *largest block palindrome*. We also show that the expected width of a largest block palindrome tends to a constant. Finally we conclude with some results on another extremal variation of block palindromes, the *smallest block palindrome*.

In Chapter 7 we present new and improved bounds on the number of length-n closed and privileged words over a k-letter alphabet. These improved bounds are the two main results of this thesis and they can be found in [42].[1]

In Chapter 8 we work with a generalization of bordered words to pairs of words. The main result of this chapter a characterization and enumeration result for this generalization of bordered words to multiple dimensions. We also give results similar to those of Nielsen [80] for this class of pairs of words. This work appears in [43].

Finally in Chapter 9 we conclude by summarizing the results of this thesis and presenting avenues for future research.

The work in Chapter 3 appears in a joint paper [27] with Trevor Clokie and Jeffrey Shallit. The work in Chapters 4, 7, and 8 appear in single-author papers [44, 42, 43].

[1] Submitted manuscript.

Chapter 2

Background

2.1 Words

Let Σ denote a finite non-empty set called an *alphabet*. The elements of Σ are called *symbols* or *letters*. Let ϵ denote the empty word. Let Σ^n denote the set of all length-n words over the alphabet Σ. We use Σ^* to denote the set of all finite words over the alphabet Σ. We write $\Sigma^+ = \Sigma^* - \{\epsilon\}$. In this thesis we primarily work with the k-letter alphabet $\Sigma_k = \{0, 1, \ldots, k-1\}$.

Let $w \in \Sigma^*$. Finite words are typically indexed starting at 1, but in this thesis we index starting at 0. If the characters of a word w are not defined explicitly (e.g., $w = w_0 w_1 \cdots w_{n-1}$), then the i'th symbol of w is denoted by $w[i]$. The length of w is denoted by $|w|$. A word x is a *subword* (or a *factor*) of w if there exist words u, v such that $w = uxv$. The subword x is a *prefix* (resp., *suffix*) of w if $u = \epsilon$ (resp., $v = \epsilon$). A word u is said to be a *proper* subword of w if $u \neq w$. The word "proper" can be used as a modifier for any type of subword, like a prefix or a suffix. For example, the word `sub` is a proper prefix of the word `subword`, but `subword` is not a proper prefix of `subword`. A word u is said to be an *internal* subword of w if $w = xuy$ for some non-empty words x and y. For example, the word `rest` is not an internal subword of the word `restoration`, but the word `rat` is.

In the combinatorics on words community, subword sometimes means subsequence, a word that can be derived from the original word by deleting 0 or more characters and preserving the same order of the remaining elements. For example, the word `sword` is a subsequence of the word `subword`. But this alternative is not studied in this thesis.

2.2 Powers

A word w is said to be a *power* if it can be written as $w = z^i$ for some non-empty word z where $i \geq 2$. Otherwise w is said to be *primitive*. For example, $\texttt{hotshots} = (\texttt{hots})^2$ is a power, but `hots` is primitive. Let $p_k(n)$ denote the number of length-n powers over Σ_k. Let $\psi_k(n)$ denote the number of length-n primitive words over Σ_k. We clearly have

$$p_k(n) = k^n - \psi_k(n).$$

From Lothaire's 1983 book [72, p. 9] we also have that

$$\psi_k(n) = \sum_{d|n} \mu(d) k^{n/d}$$

where μ is the Möbius function. The Möbius function $\mu : \mathbb{N} : \{-1, 0, 1\}$ is defined as follows:

$$\mu(n) = \begin{cases} 1, & \text{if } a^2 \nmid n \text{ for all } a > 1 \text{ and } n \text{ has an even number of prime factors;} \\ -1, & \text{if } a^2 \nmid n \text{ for all } a > 1 \text{ and } n \text{ has an odd number of prime factors;} \\ 0, & \text{if } a^2 | n \text{ for some } a > 1. \end{cases}$$

The words u and v are said to be *conjugates* of each other if there exist non-empty words x, y such that $u = xy$ and $v = yx$. If x and y are both non-empty, then v is said to be a *non-trivial* conjugate of u. If $xy = yx$, then x and y are said to *commute*. Let σ be the left-shift map, so that $\sigma^i(u) = yx$ where $u = xy$ and $|x| = i$, where i is an integer with $0 \leq i \leq |u|$. For example, any two of the words `eat`, `tea`, and `ate` are conjugates because $\texttt{eat} = \sigma(\texttt{tea}) = \sigma^2(\texttt{ate})$.

Lyndon and Schützenberger [73] showed that there is a close connection between conjugates, commutativity, and powers. They proved the following characterizations.

Theorem 2 (Lyndon-Schützenberger [73])**.** *Two non-empty words x and y commute if and only if there exists a word z, and integers $i, j \geq 1$ such that $x = z^i$ and $y = z^j$.*

Corollary 3. *Let u be a non-empty word. Then u has a non-trivial conjugate v such that $u = v$ if and only if there exists a word z, and integer $i \geq 2$ such that $u = v = z^i$.*

Let u and v be two words of equal length. The *Hamming distance* $\text{ham}(u, v)$ between u and v is defined to be the number of positions where u and v differ [53]. For example, $\text{ham}(\texttt{four}, \texttt{five}) = 3$.

As we have already seen, Lyndon and Schützenberger characterized all words x, y that commute. Alternatively, they characterized all words u that have a non-trivial conjugate v such that $\mathrm{ham}(u, v) = 0$.

One might naïvely think that the smallest possible Hamming distance between xy and yx after 0 is 1, but this is incorrect. Shallit [92] showed that $\mathrm{ham}(xy, yx) \neq 1$ for any words x and y; see Lemma 4. Thus, after 0, the smallest possible Hamming distance between xy and yx is 2. If $\mathrm{ham}(xy, yx) = 2$, then we say x and y *almost commute*.

Lemma 4 (Shallit [92])**.** *Let x and y be words. Then* $\mathrm{ham}(xy, yx) \neq 1$.

The problem of characterizing and determining the number of pairs of words that almost commute is solved Chapter 4.

2.3 Periodicity

An integer p is said to be a *period* of the word $w = a_1 a_2 \cdots a_n$ if $a_i = a_{i+p}$ for all i, $1 \leq i \leq n - p$. There are other equivalent definitions of periodicity that prove useful. One such definition is the following: An integer p is said to be a *period* of the word $w = a_1 a_2 \cdots a_n$ if there exists a length-p word x, a possibly empty prefix y of x, and positive integer j such that $w = x^j y$. A length-n word is said to be *periodic* if it has a period p with $p \leq n/2$. The period $p = 0$ is trivially a period of any word, so we generally disregard it in our analysis of the periods of a word. Historically the term "period" has also referred to the word that is repeated instead of its length. We only use the length definition in this thesis.

For example, applying both definitions of periodicity to the word $w = \texttt{alfalfa}$, we get that w has periods 3, and 6, and it can also be written as $(\texttt{alf})^2\texttt{a}$ or $(\texttt{alfalf})^1\texttt{a}$. The notions of borderedness and periodicity are quite closely related, as is shown in Theorem 5.

Theorem 5. *A word w has a period p if and only if it also has a border of length $|w| - p$.*

Proof. $\Longrightarrow$: Suppose w has a period p. Then $w = x^j y$ where j is a positive integer and $x = yz$ for some words y, z. Substituting x into w, we get $w = (yz)^j y$. Clearly $(yz)^{j-1} y$ is a border of w, and is of length $|w| - p$.

$\Longleftarrow$: Suppose w has a border of length $|w| - p$ for some $p > 0$. Then $w = xu = vx$ for some non-empty word x of length $|w| - p$ and some words u, v. Clearly we have $w[i] = w[i + |v|]$ for $1 \leq i \leq |w| - |v|$. So w has a period $|v| = |w| - |x| = p$. □

When considering periods, often the shortest are of interest. Let $\mathrm{sp}(w)$ denote the shortest period of the word w. The shortest period of a word is sometimes referred to as *the* period. Theorem 5 shows a connection between the shortest period of a word and the word's longest border. Since the shortest period and longest border are trivially related, it is natural to attempt relating shortest periods to shortest borders in some way.

The shortest border of a word must be unbordered, for otherwise the word would have a shorter border. One could ask the question, when is the shortest border also the shortest period? Or equivalently, when does a word have a unique (non-trivial) period, or exactly one unbordered border? One could also loosen the restriction of looking at borders, and just look at subwords. What relation is there between the shortest period and the longest unbordered subword? This question was asked in 1979 by Ehrenfeucht and Silberger [36]. More precisely, they conjectured that if the length of a word w is greater than or equal to twice the length of the longest unbordered subword, then the shortest period is equal to the length of the longest unbordered subword. This is illustrated in Conjecture 6. We use $\mathrm{lu}(w)$ to denote the length of the longest unbordered factor of w.

Conjecture 6 (Ehrenfeucht and Silberger [36])**.** Let w be a word. If $|w| \geq 2\,\mathrm{lu}(w)$, then $\mathrm{sp}(w) = \mathrm{lu}(w)$.

Conjecture 6 was quickly proven false later in 1979 by Assous and Pouzet [5] with the following counterexample:

$$w = 0^n 10^{n+1} 10^n 10^{n+2} 10^n 10^{n+1} 10^n. \tag{2.1}$$

A largest unbordered subword of w is $0^{n+2}10^n10^{n+1}1$, so $\mathrm{lu}(w) = 3n+6$. The shortest period of w is $4n+7$, and it is realized by the prefix $0^n10^{n+1}10^n10^{n+2}1$. Clearly $|w| = 7n+10$ is strictly greater than $2\,\mathrm{lu}(w) = 2(3n+6)$, and $\mathrm{lu}(w) \neq \mathrm{sp}(w)$. Though the $2\,\mathrm{lu}(w)$ bound was not quite right, it is still interesting to consider what the actual bound is, and how close one can get to it.

In 1982 Duval gave the first correct bound, outlined in Theorem 7.

Theorem 7 (Duval [34])**.** *Let w be a word. If $|w| \geq 4\,\mathrm{lu}(w) - 6$, then $\mathrm{sp}(w) = \mathrm{lu}(w)$.*

Along with Theorem 7, Duval also conjectured that if a word w has an unbordered prefix of length $\mathrm{lu}(w)$ then $|w| \geq 2\,\mathrm{lu}(w)$ is a sufficient condition for $\mathrm{sp}(w) = \mathrm{lu}(w)$. If Duval's conjecture were true, then it would imply that any word w has $\mathrm{sp}(w) = \mathrm{lu}(w)$ whenever $w \geq 3\,\mathrm{lu}(w)$ (see Conjecture 8).

Conjecture 8 (Duval [34])**.** Let w be a word. If $|w| \geq 3\,\mathrm{lu}(w)$, then $\mathrm{lu}(w) = \mathrm{sp}(w)$.

In 2004 Harju and Nowotka [57, 59] solved a slightly stronger version of Duval's conjecture (see Theorem 9), which is known as the "extended Duval conjecture." Holub [61] later gave an alternate proof of Theorem 9.

Theorem 9 (Harju and Nowotka [59])**.** *Let w be a word. If $|w| \geq 3\,\mathrm{lu}(w) - 2$, then $\mathrm{lu}(w) = \mathrm{sp}(w)$.*

Though Duval's conjecture has been solved, the counterexample on (2.1) that disproves Ehrenfeucht and Silberger's conjecture also shows that the bound in Theorem 9 is not optimal. So Harju and Nowotka [57, 59] conjectured that the optimal bound is close to $(7/3)\,\mathrm{lu}(w)$. Finally, Holub and Nowotka [64] resolved this conjecture, culminating in Theorem 10.

Theorem 10 (Holub and Nowotka [64])**.** *Let w be a word. If $|w| \geq \frac{7}{3}\,\mathrm{lu}(w) - 2$, then $\mathrm{sp}(w) = \mathrm{lu}(w)$.*

Moving on from the Ehrenfeucht-Silberger problem and Duval's conjecture, we turn back to Theorem 5, which connects borders to periods. Lyndon and Schützenberger [73] gave a more detailed characterization of words that have a border. They showed that if a word is bordered it must be periodic and have a certain structure.

Theorem 11 (Lyndon-Schützenberger [73])**.** *Let y be a possibly empty word and x, z be non-empty words. Then $xy = yz$ if and only if there exist words u, v, and an integer $e \geq 0$ such that $x = uv$, $z = vu$, and $y = (uv)^e u$.*

Corollary 12. *Let w be a non-empty word. Then w has a border if and only if there exists a non-empty word u, a possibly empty word v, and an integer $i \geq 1$ such that $w = (uv)^i u$.*

Recall back to Section 2.2 the definition of a power. If a word is a power then it is also periodic. We saw that Lyndon and Schützenberger characterized powers in terms of commuting words. Shallit [92] extended the notion of commutativity to allow for character mismatches. He considered $\mathrm{ham}(xy, yx)$. When $\mathrm{ham}(xy, yx) = 0$ (resp., $\mathrm{ham}(xy, yx) = 2$), we say that x and y commute (resp., almost commute).

Just as powers can be seen as a special case of periodicity, almost commuting pairs of words can be seen as a special case of a concept introduced by Klavžar and Shpectorov [68], the 2-*error border*. A word w is said to have a 2-*error border* of length i if there exists a length-i prefix u of w, and a length-i suffix u' of w such that $w = ux = yu'$ and $\mathrm{ham}(u, u') = 2$ for some x, y. The 2-error border was originally introduced in an attempt

to construct graphs that have properties similar to n-dimensional hypercubes. The n-dimensional hypercube is a graph that models Hamming distance between length-n binary words. See [100, 101] for a characterization of 2-error borders. See [13] for an algorithm to detect whether a word has a 2-error border in linear time.

2.4 Enumeration

Let $\mathcal{U}_n^k$ denote the set of length-n unbordered words over the alphabet Σ_k. Let $u_n = |\mathcal{U}_n^k|$. In 1973 Nielsen [80] studied unbordered words under the name "bifix-free words." A *bifix* of a word w is a non-empty word that is both a proper prefix and suffix of w. A word being called bifix-free means that it has no bifix, or in other words, that it is unbordered. Nielsen's main result of that 1973 paper was that the sequence u_n obeys the recurrence

$$u_n = \begin{cases} 1, & \text{if } n = 0; \\ ku_{n-1} - u_{n/2}, & \text{if } n \text{ is even}; \\ ku_{n-1}, & \text{if } n \text{ is odd}. \end{cases}$$

Nielsen also gave a procedure to efficiently list all length-n unbordered words. Additionally, he proved that the limit $\lim_{n\to\infty} u_n/k^n$ exists. In particular, he showed that for $k = 2$ there are $(c + o(1)) \cdot 2^n$ unbordered binary words, where $c \approx 0.267786$.

Later, in 1995, Lossers and Chapman [20] proved a recurrence for the number b_n of length-n bordered words over Σ_k. They showed that

$$b_n = \begin{cases} 0, & \text{if } n = 1; \\ b_{n-2} + (1 - b_{n/2})k^{-n/2}, & \text{if } n \text{ is even}; \\ b_{n-1}, & \text{if } n \text{ is odd}. \end{cases}$$

The *autocorrelation* [52, 51] of a word w, roughly speaking, is a binary word that says what borders are in w. More precisely, the *autocorrelation* of a length-n word w is $\mathrm{ac}(w) = a_1 a_2 \cdots a_n$ where

$$a_i = \begin{cases} 1, & \text{if } w \text{ has a length-}i \text{ prefix that is also a suffix of } w; \\ 0, & \text{otherwise}. \end{cases}$$

Notice that a_n is always 1 since w is a length-n prefix and suffix of itself.

Example 13. Consider the word 1011011011. It has autocorrelation ac(1011011011) = 1001001001 since its only borders are of length 1, 4, and 7.

It is quite easy to see that not every binary word is "realizable" as a correlation of some word. To see this, suppose a word w of length $n \geq 3$ has autocorrelation $0^{n-2}11$. Clearly w then has exactly one border of length $> n/2$. So w can be written as $w = ut = su$ where $u > n/2$. Then u must be bordered, and so w has more than one border.

So which binary words are autocorrelations of other words? How many valid autocorrelations are there? How many length-n words are there with a fixed autocorrelation? In 1978 Guibas and Odlyzko [51] answered these questions. They characterized all valid autocorrelations by giving necessary and sufficient conditions for an autocorrelation to be valid. They showed that there are $n^{\Theta(\log n)}$ different valid autocorrelations. They also gave a recurrence for the number of length-n words having a fixed autocorrelation.

Framing Nielsen's result differently, one could say that in a sense, he was counting the number of words with a maximum "border" of length 0. This particular framing raises some very natural questions. Is there a simple formula for the number of words of length n that have a maximum border of length t for $t \geq 1$? What is the expected length of the maximum border of a word? What is it asymptotic to? Does it converge to a constant? In 2016 Holub and Shallit [65] answered some of these questions. They gave a recurrence for the number of length-n words having a maximum border of length 1 (see Theorem 14). They also showed that over a k-letter alphabet, the expected length of the maximum border α_K of a word is asymptotic to a constant. In particular, when $k = 2$ the expected length of the maximum border α_2 of a length-n word is $1.641 + o(1)$.

Theorem 14 (Holub and Shallit [65])**.** *Let v_n denote the number of length-n words over a k-letter alphabet that have a maximum border of length 1. Then the sequence v_n obeys the recurrence*

$$v_n = \begin{cases} 0, & \text{if } n = 1; \\ k, & \text{if } n = 2; \\ kv_{n-1} - (k-1)v_{n/2}, & \text{if } n \geq 4 \text{ is even;} \\ kv_{n-1} - v_{(n+1)/2}, & \text{if } n \geq 3 \text{ is odd.} \end{cases}$$

Since Nielsen's paper there have been multiple generalizations and variations of bordered words, and with them, there have been more interesting enumeration results. Anselmo et al. [4] introduced a generalization of unbordered words to two dimensions called *unbordered pictures*. A *picture* can be thought of as a rectangular $m \times n$ matrix with values

taken from some finite alphabet. A picture p is said to be a *bordered picture* if there exists another picture p' that occurs in opposing corners of p. Otherwise p is said to be an *unbordered picture*. For example, see Example 15.

Example 15. The picture

$$\begin{bmatrix} 0 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

is bordered with borders

$$\begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \text{ and } \begin{bmatrix} 1 \end{bmatrix}.$$

In their paper, Anselmo et al. attempted to enumerate and generate unbordered pictures using similar arguments to Nielsen's. But they noted that this is not possible since Nielsen's arguments relied on a result that does not map on the the two-dimensional case. It remains an open problem to count the number of unbordered pictures of size $m \times n$ over a k-letter alphabet.

In the case of unbordered pictures, generalizing to two dimensions meant adding another co-ordinate to the words themselves so that when you refer to a symbol you need to give an x-value and a y-value (in the graphical sense). But this is not the only way to generalize unbordered words to two dimensions. One could explore a certain notion of borderedness with respect to ordered pairs of words (u, v). The two words in the pair correspond to the two dimensions.

Let u and v be words of length m and n, respectively. Let w be a non-empty word. In this thesis (u, v) to refers to an ordered pair of words. The pair (u, v) has a *right-border* if u has a non-empty proper suffix that is a prefix of v. If w is a suffix of u and prefix of v then w is said to be a *right-border* of (u, v). Analogously, the pair (u, v) has a *left-border* if u has a non-empty proper prefix that is a suffix of v. If w is a prefix of u and suffix of v then w is said to be a *left-border* of (u, v)

A pair of words (u, v) is said to be *mutually bordered* if (u, v) has both a right-border and a left-border. If (u, v) has neither a right-border nor a left-border, then (u, v) is said to be *mutually unbordered* (or *cross-bifix-free* [8]). The pair (u, v) is said to be *right-bordered* if (u, v) has a right-border but not a left-border. Similarly (u, v) is said to be *left-bordered* if (u, v) has a left-border but not a right-border.

Example 16. The pair of English words (`delivered`, `redeliver`) is mutually bordered. The word `red` is a right-border of the pair and `deliver` is a left-border of the pair.

The pair of English words (`mail`, `box`) is mutually unbordered since it has no right-border or left-border.

The pair of English words (`overlap`, `lapse`) is right-bordered. The word `lap` is a right-border of the pair.

Mutually unbordered words have previously arisen in digital communications as a generalization of a method of frame synchronization [8, 30]. The goal of frame synchronization is to let the receiver of some piece of data know where the boundaries of the frames in the data are (i.e., both the sender and receiver are on the same page). Typically this is done by inserting a specially chosen word periodically into the data stream as a kind of delimiter.

In 2000 van Wijngaarden and Willink [30] proposed a new method of frame synchronization where a set of different words are interleaved into the data stream periodically instead of appearing as a contiguous subword. An important part of frame synchronization is the detection of the periodically inserted word. In 2004 Bajic and Stojanovic [8] calculated statistical quantities related to the detection of distributed sequences in random data. For more work on mutually unbordered words, also called *cross-bifix-free* words, *mutually uncorrelated* words, or *non-overlapping* words, see [7, 12, 9, 17, 18, 71, 97, 19, 26, 96]. Also see [55, 54], where *overlap-free languages* are studied. A language is said to be *overlap-free* if no prefix of a word in the language is a suffix of another word in the language. In this thesis we choose not to use the terminology "cross-bifix-free", "mutually uncorrelated", or "non-overlapping" since we require a specific name for pairs of words (u, v) with the property that (u, v) has either a right-border or a left-border but not both.

A word is said to be *periodic* if its least period is less than or equal to half the length of the word. In some applications, such as DNA sequencing, one is working with long words that have long least periods. Keeping this in mind, one may want to extend the notion of periodicity to one that includes ordinary periodic words and also includes words that share characteristics with periodic words. In 2001 Carpi and de Luca [25] defined a class of words called *periodic-like words*, later called *closed words* [24, 6].

There are a few equivalent definition of closed words. A *complete return* [48] to the factor u in a word w is any factor of w having exactly two occurrences of u, one as a prefix and one as a suffix. Therefore a word w is closed if and only if it is a complete return to one of its factors [23, 6]. The following definition displays the connection between closed words and bordered words. A word w is said to be *closed* (or *periodic-like*) if $|w| \leq 1$ or if w has a border that occurs exactly twice in w. If u is a border w and u occurs in w exactly twice, then we say w is *closed by* u. If u is a border of w and u occurs exactly twice in w,

then w is said to be *closed by* u. It is easy to see that if a word w is closed by a word u, then u must be the largest border in w. Otherwise u would occur more than two times in w.

A word w is said to be *privileged* if $|w| \leq 1$ or if w is closed by a privileged word. See Example 17 for examples of words that are closed and privileged and words that are not closed and not privileged.

Example 17.
The English word `entanglement` has the border `ent` and only contains two occurrences of `ent`. Thus `entanglement` is a closed word, closed by `ent`. Since $|\texttt{ent}| > 1$ and `ent` is unbordered and therefore not privileged, we have that `entanglement` is not privileged.

The English word `abracadabra` is closed by `abra`. Furthermore `abra` is closed by `a`. But $|\texttt{a}| \leq 1$, so `abra` is privileged and therefore so is `abracadabra`.

The only border of the English word `eerie` is `e` and `e` appears 3 times in the word. Thus `eerie` is neither closed nor privileged.

Privileged words were first introduced in 2013 by Kellendonk et al. [67]. Later in 2013 Peltomäki [81] established some results on the basic properties of privileged words. Clearly every privileged word is also a closed word, so a lower bound for the number of privileged words also acts as a lower bound for the number of closed words. Let $C_k(n)$ denote the number of length-n closed words over a k-letter alphabet. Let $P_k(n)$ denote the number of length-n privileged words over a k-letter alphabet.

- Forsyth et al. [40] showed that $P_2(n) \geq 2^{n-5}/n^2$ for all $n > 0$.
- Nicholson and Rampersad [79] improved and generalized this bound by showing that there are constants c and n_0 such that $P_k(n) \geq c\frac{k^n}{n(\log_k(n))^2}$ for all $n \geq n_0$.
- Rukavicka [86] showed that there is a constant c such that $C_k(n) \leq c \ln n \frac{k^n}{\sqrt{n}}$ for all $n > 1$.
- Rukavicka [87] also showed that for every $j \geq 3$, there exist constants α_j and n_j such that $P_k(n) \leq \alpha_j \frac{k^n \sqrt{\ln n}}{\sqrt{n}} \ln^{\circ j}(n) \prod_{i=2}^{j-1} \sqrt{\ln^{\circ i}(n)}$ length-n privileged words for all $n \geq n_j$ where $\ln^{\circ 0}(n) = n$ and $\ln^{\circ j}(n) = \ln(\ln^{\circ j-1}(n))$.
- We improve on these bounds on the number of privileged and closed words in Chapter 7.

Closed words were first introduced in 1960 in the context of coding theory [47]. They arose as a response to a difficulty in the use of "comma-free codes." A set S of length-n words is said to be a *comma-free code* if $a_i a_{i+1} \cdots a_n b_1 b_2 \cdots b_{i-1}$ is not in S for all i with $2 \leq i \leq n$ where $(a_1 a_2 \cdots a_n,\ b_1 b_2 \cdots b_n)$ is a pair of words in S. When using a comma-free code to communicate information, the receiver can always synchronize by looking at the stream of data and checking for a codeword. A difficulty that arises with comma-free codes is that figuring out whether a given block of the stream is a codeword can be quite complicated. So Gilbert [47] defined a subset of comma-free codes (called a *prefix-synchronized code* [50]) that gets around this issue.

A code is said to be *prefix-synchronized* if every codeword starts a fixed length-p prefix $u = a_1 a_2 \cdots a_p$, and for every codeword $b_1 b_2 \cdots b_n$, the prefix u does not exist as a subword of $b_2 \cdots b_n a_1 a_2 \cdots a_{p-1} = a_2 \cdots a_p b_{p+1} \cdots b_{n-p} a_1 \cdots a_{p-1}$. Clearly the codewords of a prefix-synchronized code are not themselves closed words, but it is obvious from the definition that there is a connection. It follows from the definition of prefix-synchronized codes that for any codeword $b_1 b_2 \cdots b_n$ of such a code with common prefix of length p, the word $b_1 b_2 \cdots b_n b_1 \cdots b_p$ is closed. Conversely, if $c_1 c_2 \cdots c_n$ is a length-n closed word with a length-p border that does not appear internally, then the word $c_1 c_2 \cdots c_{n-p}$ is a codeword of a prefix-synchronized code.

Let $G_u(n)$ denote the number of length-$(n+p)$ words w such that u is a border of w and u is not an internal subword of w. Gilbert [47] conjectured that in the binary case, the prefixes u that maximize $G_u(n)$ are those of the form 1^*0. In a paper written in 1978, Guibas and Odlyzko [50] answered Gilbert's conjecture. They showed that Gilbert's conjecture is true (for large n) over the binary alphabet, and alphabets of size $k = 3$ and $k = 4$, but is not true for alphabets of size $k \geq 5$. See [77, 76] for constructions of prefix-synchronized codes.

In their seminal 1977 paper Knuth, Morris, and Pratt [69] defined a class of words called *palstars*. In order to define palstars, some other terms and notation need to be defined first. The reversal of a word w is denoted by w^R. A word w is called a *palindrome* if $w = w^R$. A word w is called a *palstar* if it can be written as a concatenation of one or more even-length palindromes. A word v is a *prime palstar* if it is a palstar but cannot be written as the concatenation of two even-length palindromes. Let $\mathrm{PS}_k(n)$ denote the number of length-n palstars over a k-letter alphabet. Let $\mathrm{PP}_k(n)$ denote the number of length-n prime palstars over a k-letter alphabet. Clearly $\mathrm{PP}_k(2m+1) = 0$ and $\mathrm{PS}_k(2m+1) = 0$ for all $m > 0$ since palstars are defined to be of even length. In 2011 Rampersad et al. [82] showed that the set of all length-$(2n)$ prime palstars is in bijection with the set of length-n unbordered words. In other words, they showed that $\mathrm{PP}_k(2n) = u_n$. In 2014 Richmond and Shallit [85]

counted the palstars by demonstrating that $\mathrm{PS}_k(n)$ obeys the recurrence

$$\mathrm{PS}_k(2n) = \sum_{i=1}^{2n} u_i \, \mathrm{PS}_k(2n-i).$$

They also showed that $\mathrm{PS}_k(2n) \in \Theta(\alpha_k^n)$ where α_k is a positive real constant with $2k-1 < \alpha_k < 2k-1/2$.

2.4.1 Conjugates

Recall that two words x and y are said to be *conjugates* of each other if there exist words u, v such that $x = uv$ and $y = vu$. Let $\sigma : \Sigma_k^* \to \Sigma_k^*$ denote the *conjugate function*, where $\sigma(\epsilon) = \epsilon$, $\sigma(cw) = wc$ for $w \in \Sigma_k^*$ and $c \in \Sigma_k$. Let $\sigma^0(w) = w$ and $\sigma^i(w) = \sigma^{i-1}(\sigma(w))$ for $i \geq 1$. For example, the word $\texttt{hotshots} = \sigma^4(\texttt{hotshots}) = (\texttt{hots})^2$ is a power and the word `hots` is primitive.

Clearly every power is bordered, and even more is true, every conjugate of a power is bordered (see Theorem 18). But not all bordered words are powers. There are $\Theta(k^n)$ bordered words of length n over a k-letter alphabet [80]. It is also easy to prove that there are $O(k^{n/2})$ powers of length n over a k-letter alphabet. Since there are so many more bordered words than there are powers, it is very natural to ask, how many unbordered conjugates must a primitive word have?

Theorem 18. *Let $w = x^i$ for some word x and some integer $i > 1$. Then $\sigma^j(w)$ is bordered for all $j \geq 0$.*

Proof. It is sufficient to show that all conjugates of w are bordered.

Let v a conjugate of w. Then there exist words st such that $v = st$ and $w = ts$. Since $w = x^i$ it follows that $t = x^j r$ and $s = zx^k$ where $i - 1 = j + k$ and $x = rz$. Then $v = st = zx^k x^j r = z(rz)^k (rz)^j r = (zr)^i$, which is clearly bordered. □

In 1971 Silberger [94] showed that every primitive word must have at least one unbordered conjugate. Silberger also conjectured that every word with k distinct letters in its representation has at least k unbordered conjugates. In 1979 Ehrenfeucht and Silberger [36] solved the conjecture, proving that every primitive word with k letters has at least k unbordered conjugates.

Theorem 19 (Ehrenfeucht and Silberger [36])**.** *Let w be a primitive word. Let a be a letter that appears in w. Then there is an unbordered conjugate of w that begins with a.*

In 2017, Holub and Müller [62] characterized all binary (primitive) words with exactly two unbordered conjugates. They showed that any binary word with exactly two unbordered conjugates can be expressed as the concatenation of two palindromes.

Let us now shift our attention to the unbordered conjugates of ordinary words. In two papers, Harju and Nowotka [56, 60] established results on the unbordered conjugates of words. In their 2004 paper [56], Harju and Nowotka focused on words over the binary alphabet. They showed that over the binary alphabet, every unbordered conjugate of a word must be followed by a bordered conjugate (see Theorem 20). In other words $\mathrm{nuc}(w) \leq \lfloor n/2 \rfloor$ for all length-n binary words w.

Theorem 20 (Harju and Nowotka [56])**.** *Let w be a binary word. Then $\sigma^i(w)$ or $\sigma^{i+1}(w)$ is bordered for all $i \geq 0$.*

Corollary 21 (Harju and Nowotka [56])**.** *Let w be a binary word of length $n > 1$. Then* $\mathrm{nuc}(w) \leq \lfloor n/2 \rfloor$.

They also showed that the only binary words w of even length that reach $\mathrm{nuc}(w) = |w|/2$ are the "cyclically overlap-free words", which only exist as words of length 2^i or $3 \cdot 2^i$ for some $i \geq 1$.

Theorem 22 (Harju and Nowotka [56])**.** *Let w be a word of length $2n$ where $n > 1$. If w has n unbordered conjugates, then it is of length 2^i or $3 \cdot 2^i$ for some $i \geq 1$.*

Let $\mathrm{mnuc}_k(n)$ denote the maximum number of unbordered conjugates of a length-n word over a k-letter alphabet. Combining Corollary 21 and Theorem 22 it follows that $\mathrm{mnuc}_2(2^i) = 2^{i-1}$ and $\mathrm{mnuc}_2(3 \cdot 2^i) = 3 \cdot 2^{i-1}$ for all $i > 1$.

This leads to a few natural questions that were left open as of Harju and Nowotka's 2004 paper, and that we answer in Chapter 3:

1. For $m > 0$, is $\mathrm{mnuc}_2(2m + 1) = m$?
2. Let $2m$ be a positive integer not of the form 2^i or $3 \cdot 2^i$ for any $i \geq 1$. Then is $\mathrm{mnuc}_2(2m) = m - 1$?
3. For any i up to $\mathrm{mnuc}_2(n)$, does there exist a binary word of length n such that it has exactly i unbordered conjugates?

In their 2008 paper, Harju and Nowotka [60] studied the unbordered conjugates of words over an alphabet of size $k \geq 3$. They showed that for any $n > 1$ and any $2 \leq i \leq n$, there exists a length-n word w over a k-letter alphabet having $\text{nuc}(w) = i$. In fact, they proved more than this, but to precisely state their results some notation is needed.

Let $\beta : \Sigma_k^* \to \Sigma_k^*$ be the *border correlation function* of a word, and defined as follows: $\beta(w) = a_0 a_1 \cdots a_{n-1}$, where

$$a_i = \begin{cases} u, & \text{if } \sigma^i(w) \text{ is unbordered;} \\ b, & \text{if } \sigma^i(w) \text{ is bordered.} \end{cases}$$

For example, $\beta(1001) = bubu$ since 1001 is bordered, 0011 is unbordered, 0110 is unbordered, and 1100 is unbordered.

Harju and Nowotka showed that for n sufficiently large one can pick any border correlation pattern v of length n except any conjugate of ub^{n-1} and there is always a word w of length n over an alphabet of size 3 (and thus a word over an alphabet of size $k \geq 3$) with $\beta(w) = v$ (see Theorem 23).

Theorem 23 (Harju and Nowotka [60])**.** *Let $n > 1$ and $k \geq 3$ be integers. Let A denote the set of all length-n words over the alphabet $\{u, b\}$.*

- *If $n \notin \{5, 7, 9, 10, 14, 17\}$, then for any $v \in A$ such that $v \notin \{\sigma^i(ub^{n-1}) : i \geq 0\}$ there exists a length-n word w over over an alphabet of size k such that $\beta(w) = v$.*
- *If $n \in \{5, 7, 9, 10, 14, 17\}$, then for any $v \neq u^n \in A$ such that $v \notin \{\sigma^i(ub^{n-1}) : i \geq 0\}$ there exists a length-n word w over an alphabet of size k such that $\beta(w) = v$.*

Corollary 24. *Let w be word of length $n > 1$ over an alphabet of size ≥ 3. Then* $\text{nuc}(w) \leq n$.

2.5 Applications

From Section 2.3 we see that borders and periods are very closely related. Borders and periodicity occur communication theory, coding theory, automata theory, and algorithms. Let us now discuss some algorithms that use borders.

The naïve way to search for a string u of length m in a larger string w of length n is to try to match u against every length-m subword of w. This strategy leads to a $O(mn)$ runtime. But there are ways to improve the runtime of string searching by using borders.

Knuth, Morris, and Pratt (KMP) [69] came up with a string searching algorithm that runs in linear time, barring any preprocessing. Their algorithm essentially uses an automaton to search for matches in the data. Suppose Σ is some finite alphabet. Let w be a word of length n and u be a word of length $m < n$. We describe the automaton for u. Each state in the KMP automaton is a prefix of u. For the purposes of this automaton, prefixes can be empty. The start state of the KMP automaton is the empty string ϵ. Suppose u' is a prefix of u and c is the character right after u' in u. If c does not exist, then transition to the state ϵ on any input. Suppose c exists. The KMP automaton will always have a transition from state u' to state $u'c$ on input c. If u' is bordered, then on input $\Sigma - \{c\}$ KMP automaton always has a transition from state u' to the state representing the largest border of u'. If u' is unbordered, then on input $\Sigma - \{c\}$ the automaton transitions to the start state ϵ. See Figure for an example of the KMP automaton.

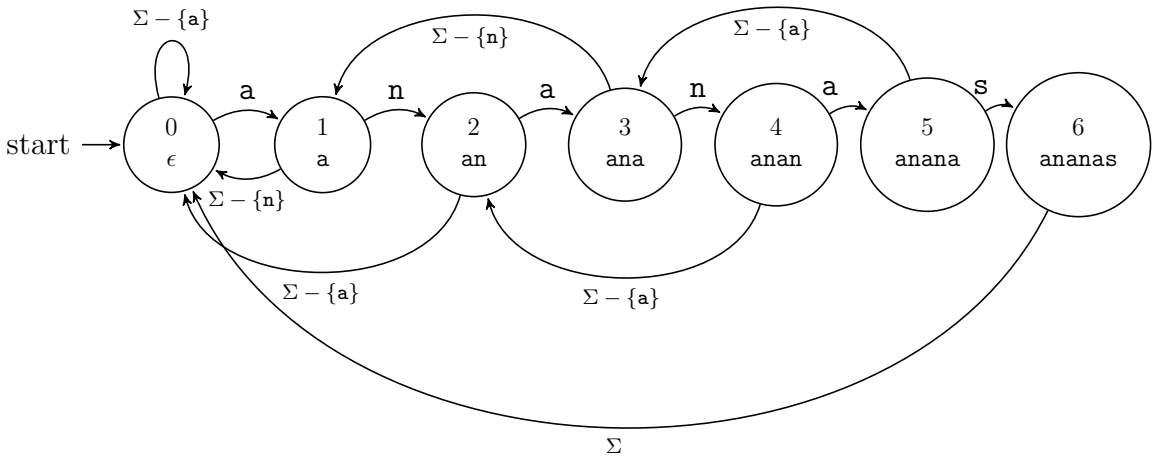

Figure 2.1: KMP automaton for the word `ananas`.

Boyer and Moore [21] also created a string searching algorithm in 1977. Note that Boyer and Moore's original paper had an error. A complete proof of the Boyer-Moore algorithm was given by Rytter [88] in 1980. Though the worst case runtime of the Boyer-Moore algorithm is still $O(mn)$, the best case is much better at $O(n/m)$. Their algorithm works by combining two heuristics, one of which is based on borders. One heuristic they use is called the "good suffix rule." The Boyer-Moore algorithm starts matching from right to left instead of left to right. Again suppose w is a word of length n and u is a word of length $m < n$. When searching for u in w with Boyer-Moore, the good suffix rule is used when some but not all characters in a suffix of u have been matched. Call this suffix that has

been matched v. When v has been matched the algorithm shifts the pattern u to the right until one of three things happens:

1. One finds another occurrence of v in u (shifting to a position where the suffix of u has border v).
2. One finds a prefix of u matches a suffix of v (shifting to a border of u).
3. The pattern u moves past the matched portion v.

Chapter 3

Unbordered conjugates

3.1 Introduction

In two fundamental papers, Harju and Nowotka [56, 60] studied the unbordered conjugates of a word. In particular, letting $\mathrm{nuc}(w)$ denote the number of unbordered conjugates of w, and $\mathrm{mnuc}_k(n)$ denote the maximum number of unbordered conjugates of a length-n word over a k-letter alphabet, they proved that

(a) for binary words w of length $n \geq 4$ we have $\mathrm{nuc}(w) \leq n/2$;

(b) for $n > 2$ even, there exists a binary word of length n having $n/2$ unbordered conjugates iff $n = 2^k$ or $n = 3 \cdot 2^k$ for some $k \geq 1$.

In other words, they explicitly computed $\mathrm{mnuc}_2(n)$ for all even n and bounded it above for odd n.

In this chapter we complete the understanding of $\mathrm{mnuc}_2(n)$ by proving that $\mathrm{mnuc}_2(n) = \lfloor n/2 \rfloor$ for all odd $n > 3$; see Theorem 25. We also show that for every possible number, up to $\mathrm{mnuc}_2(n)$, there exists a word having that number of unbordered conjugates; see Theorem 29. [1]

[1] Much of this chapter was taken verbatim from the author's paper [27].

3.2 Preliminaries

Our strategy is to show, using a decision procedure, that the maximum of $\mathrm{nuc}(w)$, over all words of length n, is actually achieved by a factor of the Thue-Morse word [98, 99]. We use the theorem-proving software `Walnut` written by Hamoon Mousavi [78] that implements this decision procedure. `Walnut` is a program that evaluates the truth of first-order statements concerning claims about k-automatic sequences. The *Thue-Morse word*

$$\mathbf{t} = 0110100110010110\cdots$$

is a well-known [3] infinite 2-automatic [3] word.

Now recall some definitions from Chapter 2.

Let $\sigma : \Sigma_k^* \to \Sigma_k^*$ denote the *cyclic shift function*, where $\sigma(\epsilon) = \epsilon$, $\sigma(cw) = wc$ for $w \in \Sigma_k^*$ and $c \in \Sigma_k$. Let $\sigma^0(w) = w$ and $\sigma^i(w) = \sigma^{i-1}(\sigma(w))$ for $i \geq 1$.

Suppose w is a binary word of length n. Let $\beta : \Sigma_k^* \to \Sigma_k^*$ be the *border correlation function* of a word (introduced by Harju and Nowotka [56]), and defined as follows: $\beta(w) = a_0 a_1 \cdots a_{n-1}$, where

$$a_i = \begin{cases} u, & \text{if } \sigma^i(w) \text{ is unbordered;} \\ b, & \text{if } \sigma^i(w) \text{ is bordered.} \end{cases}$$

3.3 Main results

A result from Harju and Nowotka [56] shows that a binary word has no two consecutive cyclic shifts that are unbordered. This result immediately tells us that a binary word of length n can have at most $\lfloor n/2 \rfloor$ unbordered conjugates. For a binary word w of even length to achieve this bound, every other cyclic shift must be unbordered, or, in other words either $\beta(w) = (ub)^{|w|/2}$ or $\beta(w) = (bu)^{|w|/2}$. Harju and Nowotka [56] showed that the only words of even length that achieve this bound are the circularly overlap-free words, which are of length $3 \cdot 2^i$ and 2^i for $i \geq 1$.

Let w be a binary word. Suppose w is of even length and is not circularly overlap-free. Clearly w cannot have $|w|/2$ unbordered conjugates, but it could potentially have $|w|/2 - 1$ unbordered conjugates. Then $\beta(w) = (ub)^i b (ub)^{|w|/2-i-1} b$ for some $i \geq 0$, up to conjugation. Now suppose w is of odd length. No circularly overlap-free words exist of odd length, so it makes sense to think that w could contain a maximum of $\lfloor |w|/2 \rfloor$ unbordered conjugates. Then $\beta(w) = (ub)^{\lfloor |w|/2 \rfloor} b$, up to conjugation.

Let w be a bordered binary word. Then $w = uvu$ for some words u and v. By the *left border* of w we mean the occurrence of u that begins at position 1 of w, and by the *right border* we mean the occurrence of u that begins at position $|w| - |u| + 1$ of w.

Theorem 25. *For all $n \geq 1$, there exists a length-n factor w of the Thue-Morse word* $\mathbf{t}$ *with* $\mathrm{nuc}(w) = \mathrm{mnuc}_2(n)$. *Furthermore, such a factor is guaranteed to occur starting at a position $\leq n$ in* $\mathbf{t}$.

Proof. When $n = 1, 2, 3$ the maximum number of unbordered conjugates $\mathrm{mnuc}_2(n)$ is achieved by the words 0, 01, and 011 respectively. Specifically we have that $\mathrm{mnuc}_2(1) = 1$, $\mathrm{mnuc}_2(2) = 2$, and $\mathrm{mnuc}_2(3) = 2$. It is readily verified that each of these words occur as a factor of the Thue-Morse word at position $\leq n$.

Let w be a length-n word at position m of the Thue-Morse word. The first step is to create a first-order predicate $\mathrm{isBorder}(l, m, n)$ that asserts that a cyclic shift of w has a border of a certain length. More specifically, we want to know whether the l'th cyclic shift of w has a border of length k. There are three cases to consider.

1. When a prefix of the right border is a suffix of w and a suffix of the right border is a prefix of w. In other words, $w = yuvx$ for words u, v, x, y where $xy = u$, $|y| = l$, and $|u| = k$. This predicate is denoted by $\mathrm{isBorderC1}(k, l, m, n)$.

2. When both borders are completely contained inside of w. In other words, $w = yuux$ for words y, u, x where $|yu| = l$, and $|u| = k$. This predicate is denoted by $\mathrm{isBorderC2}(k, l, m, n)$.

3. When a prefix of the left border is a suffix of w and a suffix of the left border is a prefix of w. In other words, $w = yvux$ for words u, v, x, y where $xy = u$, $|yvu| = l$, and $|u| = k$. This predicate is denoted by $\mathrm{isBorderC3}(k, l, m, n)$.

$$\mathrm{isBorderC1}(k,l,m,n) := ((k+l>n) \Rightarrow ((\forall i(i<n-l) \Rightarrow T[m+l+i] = T[m+l-k+i]) \wedge (\forall i(i<k+l-n) \Rightarrow T[m+i] = T[m+n-k+i])))$$

$$\mathrm{isBorderC2}(k,l,m,n) := (((k+l \leq n) \wedge (l \geq k)) \Rightarrow (\forall i\,(i<k) \Rightarrow T[m+l+i] = T[m+l-k+i]))$$

$$\mathrm{isBorderC3}(k,l,m,n) := (((k+l \leq n) \wedge (l<k)) \Rightarrow ((\forall i\,(i<k-l) \Rightarrow T[m+n-k+l+i] = T[m+l+i]) \wedge (\forall i\,(i<l) \Rightarrow T[m+i] = T[m+k+i])))$$

$$\mathrm{isBorder}(k,l,m,n) := \mathrm{isBorderC1}(k,l,m,n) \wedge \mathrm{isBorderC2}(k,l,m,n) \wedge \mathrm{isBorderC3}(k,l,m,n).$$

We define the predicate isBordered(l, m, n) that asserts that the l'th cyclic shift of a length-n word at position m in the Thue-Morse word is bordered. We can create this predicate by checking whether this word has a border of size $\leq n/2$.

$$\text{isBordered}(l, m, n) := \exists i(2i \leq n \wedge i \geq 1 \wedge \text{isBorder}(i, l, m, n)).$$

Recall that when $|w|$ is odd and w has a maximum number of unbordered conjugates, we have that $\beta(w) = (ub)^{\lfloor |w|/2 \rfloor} b$, up to conjugation. So we have exactly one pair of adjacent bordered cyclic shifts, and the rest of the cyclic shifts of w alternate between bordered and unbordered. The predicate isAlternating0(l, m, n) asserts that all of the cyclic shifts of a length-n word at position m in the Thue-Morse word alternate between unbordered and bordered, except for the l'th and $l+1$'th cyclic shifts, which are both bordered.

$$\begin{aligned}&\text{isAlternating0}(l, m, n) := \\ &\forall i(((i \neq l \wedge i < n-1) \Rightarrow (\text{isBordered}(i, m, n) = \neg\, \text{isBordered}(i+1, m, n))))\wedge \\ &(((i \neq l) \wedge (i = n-1)) \Rightarrow (\text{isBordered}(n-1, m, n) = \neg\, \text{isBordered}(0, m, n))).\end{aligned}$$

Now we create a predicate hasMNUCO(m, n) that asserts that a length-n word at position m in the Thue-Morse word achieves the maximum number of unbordered conjugates.

$$\begin{aligned}&\text{hasMNUCO}(m, n) := \exists i(((i < n-1 \wedge \text{isBordered}(i, m, n) \wedge \text{isBordered}(i+1, m, n))\vee \\ &(i = n-1 \wedge \text{isBordered}(n-1, m, n) \wedge \text{isBordered}(0, m, n))) \wedge \text{isAlternating0}(i, m, n)).\end{aligned}$$

Similarly, recall that when $|w|$ is even and w has a maximum number of unbordered conjugates, we have that $\beta(w) = (ub)^i b (ub)^{|w|/2-i-1} b$ for some $i \geq 0$ or $\beta(w) = (ub)^{|w|/2}$, up to conjugation. So we have that either all of the cyclic shifts of w alternate between bordered and unbordered, or there are exactly two pairs of adjacent bordered cyclic shifts, and the rest of the cyclic shifts of w alternate between bordered and unbordered. The predicate

$$\text{isAlternatingE}(e, l, m, n)$$

asserts that all of the cyclic shifts of a length-n word at position m in the Thue-Morse word alternate between unbordered and bordered, except for the l'th, $l+1$'th, e'th, and $e+1$'th cyclic shifts, which are all bordered. Note that isAlternatingE(n, n, m, n) asserts that all of the cyclic shifts of a length n word at position m in the Thue-Morse word alternate between unbordered and bordered.

$$\begin{aligned}\text{isAlternatingE}(e, l, m, n) := (\forall i\,(((i \neq l \wedge i \neq e \wedge i < n-1) \Rightarrow (\text{isBordered}(i, m, n) \Leftrightarrow \\ \neg\, \text{isBordered}(i+1, m, n)))) \wedge (((i \neq l) \wedge (i \neq e) \wedge (i = n-1)) \Rightarrow \\ (\text{isBordered}(n-1, m, n) \Leftrightarrow \neg\, \text{isBordered}(0, m, n))))\end{aligned}$$

Now we create a predicate hasMNUCE(m, n) that asserts that a length-n word at position m in the Thue-Morse word achieves the maximum number of unbordered conjugates.

$$\begin{aligned}\text{hasMNUCE}(m,n) := (\exists i,j\,((i<j) \wedge (i<n-1 \wedge \text{isBordered}(i,m,n) \wedge \text{isBordered}(i+1,m,n)) \wedge \\ ((j=n-1 \wedge \text{isBordered}(n-1,m,n) \wedge \text{isBordered}(0,m,n)) \vee ((j<n-1) \wedge \\ \text{isBordered}(j,m,n) \wedge \text{isBordered}(j+1,m,n))) \wedge \text{isAlternatingE}(i,j,m,n))) \vee \\ \text{isAlternatingE}(n,n,m,n).\end{aligned}$$

With these predicates we can write a predicate asserting that the Thue-Morse word contains factors of every length $n > 3$ that are maximally unbordered and occur at position $\leq n$. We split the computation into cases, one for even length words, and one for odd:

$$\begin{aligned}\forall n\,((n \geq 2) &\implies (\exists i\ \text{hasMNUCE}(i, 2n)) \wedge i \leq 2n) \\ \forall n\,((n \geq 2) &\implies (\exists i\,\text{hasMNUCO}(i, 2n+1)) \wedge i \leq 2n+1),\end{aligned}$$

and `Walnut`[2] evaluates these predicates to be true. □

Thus we have that

$$\text{mnuc}_2(n) = \begin{cases} 1, & \text{if } n = 1; \\ 2, & \text{if } n = 2 \text{ or } n = 3; \\ n/2, & \text{if } n \in \{2^{i+1}, 3\cdot 2^i : i \geq 1\}; \\ n/2-1, & \text{if } n > 3 \text{ even and } n \notin \{2^i, 3\cdot 2^i : i \geq 1\}; \\ \lfloor n/2 \rfloor, & \text{if } n > 3 \text{ odd.} \end{cases}$$

3.4 More about unbordered conjugates

In this section we show that there exist binary words of length n that have exactly i unbordered conjugates where $1 < i \leq \text{mnuc}_2(n)$.

The general idea behind the proof is to pick some $i > 1$ and then pick a word w of odd length such that $\text{nuc}(w) = i$ and $\text{mnuc}_2(|w|) = i$. Furthermore we only consider such words w such that one of w's conjugates contain 000 as a factor. Then we keep adding 0's to w precisely where 000 first occurs. This keeps the number of unbordered conjugates the same. Then we can keep increasing the size of w in this way until we hit the length we want.

[2]See Appendix A or https://cs.uwaterloo.ca/~shallit/Papers/unbordered-factors.txt for these in `Walnut` code.

Lemma 26. *For $n > 4$ odd, there exists a word $w \in \Sigma_2^n$ such that* $\mathrm{nuc}(w) = \lfloor n/2 \rfloor$ *and* 000 *is a factor of some conjugate of w.*

Proof. By Theorem 25, such a word w exists as a factor of the Thue-Morse word. It is well known that the Thue-Morse word is overlap-free. So 000 cannot be a factor of such a word w. But it is possible that $w = 0u00$, or $w = 00u0$ for some word u. We can check whether this is the case for all odd $n > 4$ by modifying our predicate from the proof of Theorem 25:

$$\forall n\,((n \geq 2) \implies (\exists i\ \mathrm{hasMNUCO}(i, 2n+1)) \wedge ((T[i] = 0 \wedge T[i+1] = 0 \wedge T[2n+i] = 0) \\ \vee (T[i] = 0 \wedge T[2n-1+i] = 0 \wedge T[2n+i] = 0))),$$

which evaluates to true. □

Lemma 27. *Let $n > 4$ be odd and w be a binary word of length n such that a conjugate of w has* 000 *as a factor and* $\mathrm{nuc}(w) = \lfloor n/2 \rfloor$*. Then every conjugate of w contains at most one distinct occurrence of* 000 *as a factor.*

Proof. Suppose, contrary to what we want to prove that a conjugate of w contains at least two distinct occurrences of 000 as a factor. Call this conjugate w'.

If the two occurrences of 000 overlap, then we can write $w' = s0000t$ for some words s, t. Then the cyclic shifts $0ts000$, $00ts00$, and $0ts000$ are bordered. This means that only $\lfloor |ts|/2 \rfloor + 1$ of the remaining cyclic shifts of w can be unbordered since any unbordered cyclic shift must be followed by a bordered one. But $\lfloor |ts|/2 \rfloor + 1 = \lfloor (n-4)/2 \rfloor + 1 < \lfloor n/2 \rfloor$, so the two occurrences of 000 cannot overlap.

If the two occurrences of 000 do not overlap, then we can write $w' = s000t000$ for some words s, t where s, and t are non-empty. Then the conjugates $00t000s0$, $0t000s00$, $00s000t0$, and $0s000t00$ are bordered. By the same argument as above, of the remaining cyclic shifts, a maximum of $\lfloor |st|/2 \rfloor + 2$ of them can be unbordered. But $\lfloor |st|/2 \rfloor + 2 = \lfloor (n-6)/2) \rfloor + 2 < \lfloor n/2 \rfloor$, a contradiction. □

Lemma 28. *Let $n > 4$ be odd and w be a binary word of length n such that a conjugate w' of w has* 000 *as a prefix and* $\mathrm{nuc}(w) = \lfloor n/2 \rfloor$*. Then* $\mathrm{nuc}(w) = \mathrm{nuc}(w') = \mathrm{nuc}(0^i w')$ *for all $i \geq 0$.*

Proof. Let $i \geq 0$ be an integer. We can write $w' = 000u$ for some word u. It is clear that $0^j u 0^{i+3-j}$ is bordered for all $1 \leq j \leq i+2$. Therefore, it suffices to prove that $s000t$ is bordered if and only if $s0^{i+3}t$ is bordered where $u = ts$.

First we prove the forward direction. Suppose $s000t$ is bordered. By Lemma 27 we have that $s000t$ contains only one occurrence of 000 as a factor. So 000 is neither a prefix of $s00$ nor a suffix of $00t$. Thus, any border of $s000t$ must of length $\leq \min\{|s|, |t|\} + 2$. But such a border would also be a border of $s0^{i+3}t$.

A similar argument works for the reverse direction. Therefore $\mathrm{nuc}(w) = \mathrm{nuc}(w') = \mathrm{nuc}(0^i w')$ for all $i \geq 0$. □

Theorem 29. *Let $k \geq 2$ and $n \geq 1$ be integers. For all $1 < i \leq \mathrm{mnuc}_k(n)$ there exists $w \in \Sigma_k^n$ such that $\mathrm{nuc}(w) = i$.*

Proof. Let $C = \{5, 7, 9, 10, 14, 17\}$. For $k \geq 4$, Harju and Nowotka [60] showed that for all integers i with $1 < i \leq n$ there exists a word $w \in \Sigma_k^n$ such that $\mathrm{nuc}(w) = i$. For $k = 3$, Harju and Nowotka [60] showed that if $n \notin C$ then for all integers i with $1 < i \leq n$ there exists a word $w \in \Sigma_k^n$ such that $\mathrm{nuc}(w) = i$, and if $n \in C$ then for all integers i with $1 < i < n$ there exists a word $w \in \Sigma_k^n$ such that $\mathrm{nuc}(w) = i$.

To the best of the author's knowledge, there is no known proof of the existence of such words for $k = 2$. Suppose $k = 2$. By Theorem 25 there exists a $w \in \Sigma_2^n$ such that w is a factor of the Thue-Morse word and $\mathrm{mnuc}_2(n) = \mathrm{nuc}(w)$. So assume $i < \mathrm{mnuc}_2(n)$. By Lemma 26 there exists a binary word u of odd length m such that $\mathrm{nuc}(u) = i = \lfloor m/2 \rfloor$ and 000 is a factor of some conjugate of u. Let u' be the conjugate of u such that 000 is a prefix of u'. Lemma 28 tells us $\mathrm{nuc}(u) = \mathrm{nuc}(u') = \mathrm{nuc}(0^{n-m}u')$. Since $\mathrm{nuc}(0^{n-m}u') = i$ and $|0^{n-m}u'| = n$, we have that for all $1 < i \leq \mathrm{mnuc}_2(n)$, there exists a $w \in \Sigma_2^n$ such that $\mathrm{nuc}(w) = i$. □

Chapter 4

Words that almost and anti-commute

4.1 Introduction

Lyndon and Schützenberger [73] characterized all words x, y that commute. Alternatively, they characterized all words u that have a non-trivial conjugate v such that $\text{ham}(u, v) = 0$.

Theorem 30 (Lyndon-Schützenberger [73])**.** *Let u be a non-empty word. Then $u = xy$ has a non-trivial conjugate $v = yx$ such that $\text{ham}(xy, yx) = 0$ if and only if there exists a word z, and integers $i, j \geq 1$ such that $x = z^i$, $y = z^j$, and $u = v = z^{i+j}$.*

Later, Fine and Wilf [39] showed that one can achieve the forward implication of Theorem 30 with a weaker hypothesis. Namely, that xy and yx need not be equal, but only agree on the first $|x| + |y| - \gcd(|x|, |y|)$ terms.

Theorem 31 (Fine-Wilf [39])**.** *Let x and y be non-empty words. If xy and yx agree on a prefix of length at least $|x| + |y| - \gcd(|x|, |y|)$, then there exists a word z, and integers $i, j \geq 1$ such that $x = z^i$, $y = z^j$, and $xy = yx = z^{i+j}$.*

Fine and Wilf also showed that the bound of $|x| + |y| - \gcd(|x|, |y|)$ is optimal, in the sense that if xy and yx agree only on the first $|x| + |y| - \gcd(|x|, |y|) - 1$ terms, then xy need not equal yx. They demonstrated this by constructing words x, y of any length such that xy and yx agree on the first $|x| + |y| - \gcd(|x|, |y|) - 1$ terms and differ at position $|x| + |y| - \gcd(|x| + |y|)$. We call pairs of words x, y of this form *Fine-Wilf pairs*.

These words have been shown to have a close relationship with the well-known *finite Sturmian words* [32].

Example 32. We give some examples of words that display the optimality of the Fine-Wilf result.

Let $x = 000000010000$ and $y = 00000001$. Then $|x| = 12$, $|y| = 8$, and $\gcd(|x|, |y|) = 4$.

$$xy = 00000001000000000001$$
$$yx = 00000001000000010000$$

Let $x = 010100101010$ and $y = 0101001$. Then $|x| = 12$, $|y| = 7$, and $\gcd(|x|, |y|) = 1$.

$$xy = 01010010101001010001$$
$$yx = 01010010101001010010$$

One remarkable property of these words is that they "almost" commute, in the sense that xy and yx agree for as long a prefix as possible and differ in as few positions as possible. See Lemma 34 for a proof of this property.

One might naïvely think that the smallest possible Hamming distance between xy and yx after 0 is 1, but this is incorrect. It turns out that $\text{ham}(xy, yx) \neq 1$ for any words x and y; see Lemma 33. Thus, after 0, the smallest possible Hamming distance between xy and yx is 2. If $\text{ham}(xy, yx) = 2$, then we say x and y *almost commute*.

Lemma 33 (Shallit [92])**.** *Let x and y be words. Then* $\text{ham}(xy, yx) \neq 1$.

As we saw in Section 2.3, the concept of words almost commuting is very similar to the 2-*error border* introduced by Klavžar and Shpectorov [68].

In this chapter, we characterize and count all pairs of words x, y that almost commute. We also characterize and count all pairs of words x, y such that $\text{ham}(xy, yx) = |xy|$ (i.e., x and y *anti-commute*).

Let n and i be integers such that $n > i \geq 1$. Let $H_m(n)$ denote the set of length-n words u over Σ_k that have a conjugate v such that $\text{ham}(u, v) = m$. Let $h_m(n) = |H_m(n)|$. Let $H_m(n, i)$ denote the set of length-n words u over Σ_k such that $\text{ham}(u, \sigma^i(u)) = m$. Let $h_m(n, i) = |H_m(n, i)|$.

The rest of the chapter is structured as follows. In Section 4.2 we prove that Fine-Wilf pairs almost commute. In Section 4.3 we characterize the words in $H_2(n, i)$ and present a formula to calculate $h_2(n, i)$; see Lemma 35 and Lemma 36. In Section 4.4 we characterize the words in $H_n(n, i)$ and present a formula to calculate $h_n(n, i)$; see Lemma 38 and Corollary 39. In Section 4.5 we prove some properties of $H_m(n, i)$ and $H_m(n)$ that we

make use of in later sections. In Section 4.6 we present a formula to calculate $h_2(n)$; see Theorem 49. In Section 4.7 we count the number of length-n words u with *exactly* one conjugate such that $\mathrm{ham}(u, v) = 2$. In Section 4.8 we count the number of Lyndon words in $H_m(n)$. Finally, in Section 4.9 we show that $h_2(n)$ grows erratically.[1]

4.2 Fine-Wilf pairs almost commute

In this section we prove that Fine-Wilf pairs almost commute. This result appears without proof in [93]. This result is also basically a special case of Theorem 2.3.5 in [91].

Lemma 34. *Let x and y be non-empty words. Suppose xy and yx agree on a prefix of length $|x| + |y| - \gcd(|x|, |y|) - 1$ but disagree at position $|x| + |y| - \gcd(|x|, |y|)$. Then $\mathrm{ham}(xy, yx) = 2$.*

Proof. The proof is by induction on $|x|+|y|$. Suppose xy and yx agree on a prefix of length $|x| + |y| - \gcd(|x|, |y|) - 1$ but disagree at position $|x| + |y| - \gcd(|x|, |y|)$. Without loss of generality, let $|x| \leq |y|$.

First, we take care of the case when $|x| = |y|$, which also takes care of the base case $|x| + |y| = 2$. Since $|x| = |y|$, we have that $\gcd(|x|, |y|) = |x| = |y|$. Therefore, x and y share a prefix of length $|x| + |y| - \gcd(|x|, |y|) - 1 = |x| - 1$ but disagree at position $|x|$. This implies that $\mathrm{ham}(x, y) = 1$. Thus $\mathrm{ham}(xy, yx) = 2\,\mathrm{ham}(x, y) = 2$.

Suppose $|x| < |y|$. Then $\gcd(|x|, |y|) \leq |x|$. So $|x| + |y| - \gcd(|x|, |y|) - 1 \geq |y| - 1$. Thus xy and yx must share a prefix of length $\geq |y| - 1$. However, since $|x| < |y|$, we have that x must then be a proper prefix of y. So write $y = xt$ for some non-empty word t. Then $\mathrm{ham}(xy, yx) = \mathrm{ham}(xxt, xtx) = \mathrm{ham}(xt, tx)$. Since xt, tx are suffixes of xy, yx we have that xt and tx agree on the first $|y| - \gcd(|x|, |y|) - 1$ terms and disagree at position $|y| - \gcd(|x|, |y|)$. Clearly $\gcd(|x|, |y|) = \gcd(|x|, |xt|) = \gcd(|x|, |x| + |t|) = \gcd(|x|, |t|)$, and $|y| - \gcd(|x|, |y|) = |x| + |t| - \gcd(|x|, |t|)$. Therefore xt and tx share a prefix of length $|x| + |t| - \gcd(|x|, |t|) - 1$ and differ at position $|x| + |t| - \gcd(|x|, |t|)$. By induction $\mathrm{ham}(xt, tx) = 2$, and thus $\mathrm{ham}(xy, yx) = 2$. □

[1] This chapter is taken almost verbatim from the author's article [44].

4.3 Almost-commuting words

In this section we characterize the words in $H_2(n,i)$ and use this characterization to provide an explicit formula for $h_2(n,i)$.

Lemma 35. *Let n, i be positive integers such that $n > i$. Let $g = \gcd(n,i)$. Let w be a length-n word. Let $w = x_0x_1\cdots x_{n/g-1}$ where $|x_j| = g$ for all j, $0 \le j \le n/g-1$. Then $w \in H_2(n,i)$ iff there exist two distinct integers j_1, j_2, $0 \le j_1 < j_2 \le n/g-1$ such that $\mathrm{ham}(x_{j_1}, x_{j_2}) = 1$ and $x_j = x_{(j+i/g) \bmod n/g}$ for all $j \neq j_1, j_2$, $0 \le j \le n/g-1$.*

Proof. We write $w = x_0x_1\cdots x_{n/g-1}$ where $|x_j| = g$ for all j, $0 \le j \le n/g-1$. Since g divides i, we have that $\sigma^i(w) = x_{i/g}\cdots x_{n/g-1}x_0\cdots x_{i/g-1}$.

$\Longrightarrow$: Suppose $w \in H_2(n,i)$. Then

$$\begin{aligned}\mathrm{ham}(w,\sigma^i(w)) &= \mathrm{ham}(x_0x_1\cdots x_{n/g-1}, x_{i/g}\cdots x_{n/g-1}x_0\cdots x_{i/g-1})\\ &= \sum_{j=0}^{n/g-1} \mathrm{ham}(x_j, x_{(j+i/g) \bmod n/g})\\ &= 2.\end{aligned}$$

In order for the Hamming distance between w and $\sigma^i(w)$ to be 2, we must have that either

- $\mathrm{ham}(x_j, x_{(j+i/g) \bmod n/g}) = 2$ for exactly one j, $0 \le j \le n/g-1$; or
- $\mathrm{ham}(x_{j_1}, x_{(j_1+i/g) \bmod n/g}) = 1$ and $\mathrm{ham}(x_{j_2}, x_{(j_2+i/g) \bmod n/g}) = 1$ for two distinct integers j_1, j_2, $0 \le j_1 < j_2 \le n/g-1$.

Suppose $\mathrm{ham}(x_j, x_{(j+i/g) \bmod n/g}) = 2$ for some j, $0 \le j \le n/g-1$. Then it follows that $x_p = x_{(p+i/g) \bmod n/g}$ for all $p \neq j$, $0 \le p \le n/g-1$. Since $g = \gcd(n,i)$, we have that $\gcd(n/g, i/g) = 1$. The additive order of i/g modulo n/g is $\frac{n/g}{\gcd(n/g,i/g)} = n/g$. Therefore, we have that

$$x_{(j+i/g) \bmod n/g} = x_{(j+2i/g) \bmod n/g} = \cdots = x_{(j+(n/g-1)i/g) \bmod n/g} = x_j$$

and $\mathrm{ham}(x_j, x_{(j+i/g) \bmod n/g}) = 2$, a contradiction.

Suppose $\mathrm{ham}(x_{j_1}, x_{(j_1+i/g) \bmod n/g}) = 1$ and $\mathrm{ham}(x_{j_2}, x_{(j_2+i/g) \bmod n/g}) = 1$ for two distinct integers j_1, j_2, $0 \le j_1 < j_2 \le n/g-1$. Then it follows that $x_j = x_{(j+i/g) \bmod n/g}$ for all $j \neq j_1, j_2$, $0 \le j \le n/g-1$. Since the additive order of i/g modulo n/g is n/g, we have

that if we start at j_1 and successively add i/g and take the result modulo n/g, then we will reach every integer between 0 and $n/g-1$. Therefore, we will reach j_2 before we reach j_1 again. Thus, since $x_j = x_{(j+i/g) \bmod n/g}$ for all $j \neq j_1, j_2$, $0 \leq j \leq n/g-1$, we have that

$$x_{(j_1+i/g) \bmod n/g} = x_{(j_1+2i/g) \bmod n/g} = \cdots = x_{j_2}.$$

But now we have $\mathrm{ham}(x_{j_1}, x_{(j_1+i/g) \bmod n/g}) = 1$ and $x_{(j_1+i/g) \bmod n/g} = x_{j_2}$, which implies $\mathrm{ham}(x_{j_1}, x_{j_2}) = 1$.

$\Longleftarrow$: Suppose there exist two distinct integers j_1, j_2, $0 \leq j_1 < j_2 \leq n/g-1$ such that $\mathrm{ham}(x_{j_1}, x_{j_2}) = 1$ and $x_j = x_{(j+i/g) \bmod n/g}$ for all $j \neq j_1, j_2$, $0 \leq j \leq n/g-1$. Since the additive order of i/g modulo n/g is n/g, we have that if we start at j_1 and successively add i/g modulo n/g, then we will reach every integer between 0 and $n/g-1$. But this means that we will reach j_2 before we get to j_1 again. Thus, we have that

$$x_{(j_1+i/g) \bmod n/g} = x_{(j_1+2i/g) \bmod n/g} = \cdots = x_{j_2}.$$

Similarly, if we start at j_2 and successively add i/g modulo n/g we will reach j_1 before looping back to j_2. So

$$x_{(j_2+i/g) \bmod n/g} = x_{(j_2+2i/g) \bmod n/g} = \cdots = x_{j_1}.$$

Therefore, we have that $w \in H_2(n,i)$ since

$$\begin{aligned}
\mathrm{ham}(w, \sigma^i(w)) &= \mathrm{ham}(x_0 x_1 \cdots x_{n/g-1}, x_{i/g} \cdots x_{n/g-1} x_0 \cdots x_{i/g-1}) \\
&= \sum_{j=0}^{n/g-1} \mathrm{ham}(x_j, x_{(j+i/g) \bmod n/g}) \\
&= \mathrm{ham}(x_{j_1}, x_{(j_1+i/g) \bmod n/g}) + \mathrm{ham}(x_{j_2}, x_{(j_2+i/g) \bmod n/g}) \\
&= \mathrm{ham}(x_{j_1}, x_{j_2}) + \mathrm{ham}(x_{j_2}, x_{j_1}) \\
&= 2.
\end{aligned}$$

□

Lemma 36. *Let n, i, and k be integers such that $k \geq 2$ and $n > i \geq 1$. Then*

$$h_2(n,i) = \frac{1}{2} k^{\gcd(n,i)} (k-1) n \left(\frac{n}{\gcd(n,i)} - 1 \right).$$

Proof. Let w be a length-n word. Let $g = \gcd(n, i)$. We split up w into length-g blocks. We write $w = x_0 x_1 \cdots x_{n/g-1}$ where $|x_j| = g$ for all j, $0 \le j \le n/g - 1$. Lemma 35 gives a complete characterization of $H(n, i)$. Namely, the word w is in $H(n, i)$ if and only if there exist two distinct integers j_1, j_2, $0 \le j_1 < j_2 \le n/g - 1$ such that $\text{ham}(x_{j_1}, x_{j_2}) = 1$ and $x_j = x_{(j+i/g) \bmod n/g}$ for all $j \ne j_1, j_2$, $0 \le j \le n/g - 1$. Given j_1, j_2, x_{j_1}, and x_{j_2}, all x_j for $j \ne j_1, j_2$, $0 \le j \le n/g - 1$ are already determined.

There are

$$\sum_{j_2=1}^{n/g-1} \sum_{j_1=0}^{j_2-1} 1 = \frac{1}{2}\frac{n}{g}\left(\frac{n}{g} - 1\right)$$

choices for j_1 and j_2. There are k^g options for x_{j_1}. Considering that x_{j_1} and x_{j_2} differ in exactly one position, there are $g(k-1)$ choices for x_{j_2} given x_{j_1}. Putting everything together we have that

$$\begin{aligned} h_2(n, i) &= \overbrace{\frac{1}{2}\frac{n}{g}\left(\frac{n}{g} - 1\right)}^{\text{choices for } j_1 \text{ and } j_2} \overbrace{k^g}^{\text{choices for } x_{j_1}} \overbrace{g(k-1)}^{\text{choices for } x_{j_2} \text{ given } x_{j_1}} \\ &= \frac{1}{2} k^{\gcd(n,i)} (k-1) n \left(\frac{n}{\gcd(n, i)} - 1\right). \end{aligned}$$

□

Corollary 37. *Let $m, n \ge 1$ and $k \ge 2$ be integers. Then there are exactly*

$$h_2(n+m, m) = \frac{1}{2} k^{\gcd(n+m,m)} (k-1)(n+m) \left(\frac{n+m}{\gcd(n+m, m)} - 1\right).$$

pairs of words (x, y) of length (m, n) such that $\text{ham}(xy, yx) = 2$.

4.4 Anti-commuting words

We say that the words x and y *anti-commute* if xy and yx differ maximally, i.e., $\text{ham}(xy, yx) = |xy|$. Let w be a length-n word. Then w is in $H_n(n, i)$ iff we can write $w = uv$ where $|u| = i$ and $\text{ham}(uv, vu) = n = |uv|$ iff u and v anti-commute. Thus there are $h_{n+m}(n+m, m)$ anti-commuting pairs of words (u, v) of length (m, n). In this section we characterize and count $H_n(n, i)$, giving a formula for the number of anti-commuting pairs of words.

First we introduce some definitions. A *graph* G is an ordered pair (V, E) consisting of V, a set of *vertices* (or *nodes*), and $E \subseteq \{\{u, v\} : u, v \in V\}$, a set of unordered pairs of

vertices, called *edges*. We say that two vertices $u, v \in V$ are adjacent if $\{u, v\} \in E$. The graph $G = (V, E)$ is said to be a *length-n cycle graph* if $|V| = n$ and $E = \{\{v_i, v_{(i+1) \bmod n}\} : 0 \leq i \leq n-1\}$ where $V = \{v_i : 0 \leq i \leq n-1\}$. A k-colouring of a graph G is an assignment of integers $\{0, 1, \ldots, k-1\}$, or *colours*, to the vertices of G such that all adjacent vertices have distinct colours.

Let $w = w_0 w_1 \cdots w_{n-1}$ be a length-n word over Σ_k. Let i be a positive integer such that $i < n$. Let $f(w, i)$ denote the graph (V, E) where $V = \{w_j : 0 \leq j \leq n-1\}$ and $E = \{\{w_j, w_{(j+i) \bmod n}\} : 0 \leq j \leq n-1\}$.

Lemma 38. *Let $w = w_0 w_1 \cdots w_{n-1}$ be a length-n word over Σ_k. Let i be a positive integer with $i < n$. Let $g = \gcd(n, i)$. Then $w \in H_n(n, i)$ if and only if $f(w, i)$ is composed of g disjoint length-(n/g) k-coloured cycle graphs.*

Proof. $\Longrightarrow$: Suppose $w \in H_n(n, i)$. Then

$$\begin{aligned} \mathrm{ham}(w, \sigma^i(w)) &= \mathrm{ham}(w_0 w_1 \cdots w_{n-1}, w_i \cdots w_{n-1} w_0 \cdots w_{i-1}) \\ &= \sum_{j=0}^{n-1} \mathrm{ham}(w_j, w_{(j+i) \bmod n}) \\ &= n. \end{aligned}$$

In order for the Hamming distance between w and $\sigma^i(w)$ to be n, we must have that $\mathrm{ham}(w_j, w_{(j+i) \bmod n}) = 1$ for all j, $0 \leq j \leq n-1$. But from the definition of the edge set of $f(w, i)$ this implies that all adjacent vertices in $f(w, i)$ are labelled with distinct colours. The additive order of i modulo n is n/g, which implies that each separate vertex $w_0, w_1, \ldots, w_{g-1}$ is on a separate disjoint length-(n/g) cycle. Thus $f(w, i)$ is composed of g disjoint length-(n/g) k-coloured cycle graphs.

$\Longleftarrow$: By definition, the vertices w_j and $w_{(j+i) \bmod n}$ are adjacent in $f(w, i)$. Since $f(w, i)$ is k-coloured, we have that adjacent vertices labelled with distinct colours. Thus

$$\mathrm{ham}(w, \sigma^i(w)) = \sum_{j=0}^{n-1} \mathrm{ham}(w_j, w_{(j+i) \bmod n}) = n.$$

□

Let $C(n, k)$ be the number of valid k-colourings of a length-n cycle graph. It is well known and easy to prove that $C(1, k) = 0$, $C(2, k) = k(k-1)$, and

$$\begin{aligned} C(n, k) &= (k-2)C(n-1, k) + (k-1)C(n-2, k) \\ &= (k-1)^n + (-1)^n (k-1) \end{aligned}$$

for $n \geq 3$.

Corollary 39. *Let $k \geq 2$ and $n > i \geq 1$ be integers. Let $g = \gcd(n, i)$. Then*

$$h_n(n, i) = C(n/g, k)^g.$$

Corollary 40. *Let $k \geq 2$ and $m, n \geq 1$ be integers. Let $g = \gcd(n + m, m)$. Then there are exactly*

$$h_{n+m}(n + m, m) = C((n + m)/g, k)^g$$

pairs of words (x, y) of length (m, n) such that $\text{ham}(xy, yx) = |xy|$.

4.5 Some useful properties

In this section we prove some properties of $H_m(n, i)$ and $H_2(n)$ that we use in later sections.

Lemma 41. *Let u be a length-n word. Let i be an integer with $0 < i < n$. If $u \in H_m(n, i)$ then $u \in H_m(n, n - i)$.*

Proof. Suppose $i \leq n/2$. Then we can write $u = xtz$ for some words t, z where $|x| = |z| = i$ and $|t| = n - 2i$. We have that $\text{ham}(xtz, tzx) = \text{ham}(xt, tz) + \text{ham}(z, x) = m$. Consider the word zxt. Clearly $v = zxt$ is a conjugate of $u = xtz$ such that $\text{ham}(xtz, zxt) = \text{ham}(x, z) + \text{ham}(tz, xt) = m$ where $u = (xt)z$ and $v = z(xt)$ with $|xt| = n - i$. Therefore $u \in H_m(n, n - i)$.

Suppose $i > n/2$. Then we can write $u = zty$ for some words t, z where $|z| = |y| = n - i$ and $|t| = 2i - n$. We have that $\text{ham}(zty, yzt) = \text{ham}(z, y) + \text{ham}(ty, zt) = m$. Consider the word tyz. Clearly $v = tyz$ is a conjugate of $u = zty$ such that $\text{ham}(zty, tyz) = \text{ham}(zt, ty) + \text{ham}(y, z) = m$ where $u = z(ty)$ and $v = (ty)z$ with $|z| = n - i$. Therefore $u \in H_m(n, n - i)$. □

Lemma 42. *Let u be a length-n word. If $u \in H_2(n)$, then $\text{ham}(u, v) > 0$ for any non-trivial conjugate v of u.*

Proof. We prove the contrapositive of the lemma statement. Namely, we prove that if there exists a non-trivial conjugate v of u such that $\text{ham}(u, v) = 0$ then $u \notin H_2(n)$.

Suppose $u = xy$ and $v = yx$ for some non-empty words x, y. Then by Theorem 30 we have that there exists a word z, and an integer $i \geq 2$ such that $u = v = z^i$. Let w be a conjugate of u. Then $w = (ts)^i$ where $z = st$. So $\text{ham}(u, w) = \text{ham}((st)^i, (ts)^i) = i\,\text{ham}(st, ts)$. If $st = ts$, then $\text{ham}(u, w) = 0$. If $st \neq ts$, then $\text{ham}(st, ts) \geq 2$ (Lemma 33). Since $\text{ham}(st, ts) \geq 2$ and $i \geq 2$, we have $\text{ham}(u, w) \geq 4$. Thus $u \notin H_2(n)$. □

Corollary 43. *Let u be a length-n word. If u is a power, then $u \notin H_2(n)$.*

Corollary 44. *All words in $H_2(n)$ are primitive.*

Lemma 45. *Let u be a length-n word. Let i be an integer with $0 < i < n$. If $u \in H(n,i)$, then any conjugate of u is also in $H(n,i)$.*

Proof. Suppose $u \in H(n,i)$. Then $\text{ham}(u, \sigma^i(u)) = 2$. If we shift both u and $\sigma^i(u)$ by the same amount, then the symbols that are being compared to each other do not change. Thus $\text{ham}(\sigma^j(u), \sigma^{i+j}(u)) = 2$ for all $j \geq 0$. So any conjugate $\sigma^j(u)$ of u must also be in $H(n,i)$. □

4.6 Counting almost-commuting words

Lemma 41 shows that $H_m(n,i) = H_m(n,n-i)$, which in turn implies that $h_m(n) \leq \sum_{i=1}^{\lfloor n/2 \rfloor} h_m(n,i)$. To make this inequality an equality we need to be able to account for those words that are double-counted in the sum $\sum_{i=1}^{\lfloor n/2 \rfloor} h_m(n,i)$. In this section we resolve this problem for the case when $m = 2$ and give an exact formula for $h_2(n)$. More specifically, we show that all words w that are in both $H_2(n,i)$ and $H_2(n,j)$, for $i \neq j$, must exhibit a certain regular structure that we can explicitly describe. Then we use this structure result, in addition to the results from Section 4.3 and Section 4.5, to give an exact formula for $h_2(n)$. See A179674 in the *On-Line Encyclopedia of Integer Sequences* (OEIS) [95] for the sequence $(h_2(n))_{n \geq 0}$.

Lemma 46. *Let n, i, j be positive integers such that $n \geq 2i > 2j$. Let $g = \gcd(n,i,j)$. Let w be a length-n word. Then $w \in H_2(n,i)$ and $w \in H_2(n,j)$ if and only if there exists a word u of length g, a word v of length g with $\text{ham}(u,v) = 1$, and a non-negative integer $p < n/g$ such that $w = u^p v u^{n/g-p-1}$.*

Proof.
$\Longrightarrow$: The proof is by induction on $|w| = n$. Suppose $w \in H_2(n,i)$ and $w \in H_2(n,j)$. First, we take care of the case when $n = 2i$, which also includes the base case $n = 4$, $i = 2$, $j = 1$. Write $w = xyx'y'$ where $|xy| = |x'y'| = i = n/2$ and $|x| = |x'| = j$. Since $w \in H_2(n,i)$, we have that $\text{ham}(xyx'y', x'y'xy) = 2$. This implies that $\text{ham}(xy, x'y') = 1$. Furthermore, if $\text{ham}(xy, x'y') = 1$ then either $\text{ham}(x,x') = 1$ or $\text{ham}(y,y') = 1$.

Suppose $\text{ham}(x,x') = 1$. Then $y = y'$. Since $w \in H_2(n,j)$, we have $\text{ham}(xyx'y, yx'yx) = \text{ham}(xy, yx') + \text{ham}(x'y, yx) = 2$. Suppose $\text{ham}(xy, yx') = 0$ or $\text{ham}(x'y, yx) = 0$. Both

cases imply that $\text{ham}(xy, yx) = 1$, which contradicts Lemma 33. Thus, we must have $\text{ham}(xy, yx') = \text{ham}(x'y, yx) = 1$. But this implies that

- $\text{ham}(xy, yx) = 0$ and $\text{ham}(x'y, yx') = 2$, or
- $\text{ham}(xy, yx) = 2$ and $\text{ham}(x'y, yx') = 0$.

Without loss of generality, suppose $\text{ham}(xy, yx) = 0$. By Theorem 30, there exists a word s, and integers $l, m \geq 1$ such that $x = s^l$ and $y = s^m$. Clearly $|s|$ divides $\gcd(n/2, j) = \gcd(n, n/2, j) = \gcd(n, i, j) = g$ since it divides both $|x| = j$ and $|xy| = i = n/2$. Therefore, there exists a length-g word u such that $x = u^{j/g}$ and $y = u^{(i-j)/g}$. Since x and x' differ in exactly one position, and $x = u^{j/g}$, there exists a length-g word v with $\text{ham}(u, v) = 1$, and a non-negative integer $p' < j/g$ such that $x' = u^{p'}vu^{j/g-p'-1}$. Letting $p = p' + i/g = p' + (n/2)/g$, we have $w = xyx'y = u^{i/g}u^{p'}vu^{j/g-p'-1}u^{(i-j)/g} = u^pvu^{n/g-p-1}$.

Suppose $\text{ham}(y, y') = 1$. Then $x = x'$. Since $w \in H_2(n, j)$, we have $\text{ham}(xyxy', yxy'x) = \text{ham}(xy, yx) + \text{ham}(xy', y'x) = 2$. By Lemma 33, we have that $\text{ham}(xy, yx) \neq 1$ and $\text{ham}(xy', y'x) \neq 1$. So either $\text{ham}(xy, yx) = 0$ or $\text{ham}(xy', y'x) = 0$. Without loss of generality, suppose $\text{ham}(xy, yx) = 0$. As in the previous case when $\text{ham}(x, x') = 1$, there exists a length-g word u such that $x = u^{j/g}$ and $y = u^{(i-j)/g}$. Since y and y' differ in exactly one position, there exists a length-g word v with $\text{ham}(u, v) = 1$, and a non-negative integer $p' < (i-j)/g$ such that $y' = u^{p'}vu^{(i-j)/g-p'-1}$. Letting $p = p' + (i+j)/g = p' + (n/2+j)/g$, we have $w = xyxy' = u^{i/g}u^{j/g}u^{p'}vu^{(i-j)/g-p'-1} = u^pvu^{n/g-p-1}$.

Now, we take care of the case when $n > 2i$. Write $w = xyx'y'z$ for words x, y, x', y', z where $|xy| = |x'y'| = i$, and $|x| = |x'| = j$. Since $w \in H_2(n, i)$, we have that w and $\sigma^i(w)$ differ in exactly two positions $j_1 < j_2$. But $n > 2i$ implies that either

- $j_2 - j_1 > i$, or
- $j_2 - j_1 \leq i$ and $n - (j_2 - j_1) > 2i - (j_2 - j_1) \geq i$.

In either case we have that there is a length-i contiguous block, possibly occurring in the wraparound, where w and $\sigma^i(w)$ match. This translates to there being a length-$2i$ block in w of the form tt where $|t| = i$. Additionally, we have that $\sigma^m(w) \in H_2(n, i)$ and $\sigma^m(w) \in H_2(n, j)$ for all $m \geq 0$ by Lemma 45. Therefore, we can assume without loss of generality that w begins with this length-$2i$ block (i.e., $\text{ham}(xy, x'y') = 0$).

Suppose $\text{ham}(xy, x'y') = 0$. Then $\text{ham}(xyxyz, xyzxy) = \text{ham}(xyxyz, yxyzx) = 2$. Clearly $\text{ham}(xyxyz, xyzxy) = \text{ham}(xyz, zxy) = 2$, so $xyz \in H_2(n-i, i)$. Now, either $xy =$

yx or $xy \neq yx$. If $xy = yx$, then we clearly have $\text{ham}(xyxyz, yxyzx) = \text{ham}(xyz, yzx) = 2$. Therefore, we have $xyz \in H_2(n-i, j)$. Let $g = \gcd(n-i, i, j)$. We have that $g = \gcd(n-i, i, j) = \gcd(\gcd(n-i, i), j) = \gcd(\gcd(n, i), j) = \gcd(n, i, j)$. If $n - i \geq 2i > 2j$, then we can apply induction to xyz directly. By Lemma 41, we have that if $xyz \in H_2(n-i, i)$ and $xyz \in H_2(n-i, j)$, then $xyz \in H_2(n-i, n-2i)$ and $xyz \in H_2(n-i, n-i-j)$. If $n-i < 2i$ and $n-i \geq 2j$, then $n-i > 2(n-2i)$ and $\gcd(n-i, n-2i, j) = \gcd(n, i, j) = g$. However, in this case we can have $j = n-2i$, which we have to take care of separately since it does not satisfy the inductive hypothesis. If $n - i < 2j < 2i$, then $n - i > 2(n-i-j)$, $n - i > 2(n-2i)$, and $\gcd(n-i, n-2i, n-i-j) = \gcd(n, i, j) = g$.

Suppose $j \neq n - 2i$. By induction there exists a word u of length g, a word v of length g with $\text{ham}(u, v) = 1$, and a non-negative integer $p' < (n-i)/g$ such that $xyz = u^{p'}vu^{(n-i)/g-p'-1}$. Since $xy = yx$ and $g \mid \gcd(i, j)$, it is clear that $xy = u^{i/g}$. Then $w = xyxyz = u^{p'+i/g}vu^{(n-i)/g-p'-1}$. Letting $p = p' + i/g$, we have $w = u^pvu^{n/g-p-1}$.

Suppose $j = n-2i$. Then $w = xyxyz$ where $|z| = |x| = n-2i$. Since $w \in H_2(n, n-2i)$, we have $\text{ham}(xyxyz, yxyzx) = \text{ham}(xy, yx) + \text{ham}(xy, yz) + \text{ham}(z, x) = 2$. But $xy = yx$ by assumption. Thus $\text{ham}(xy, yz) + \text{ham}(z, x) = 2$, which is only true when $\text{ham}(z, x) = 1$. By Theorem 30, there exists a word s, and integers $l, m \geq 1$ such that $x = s^l$ and $y = s^m$. Since $|s|$ divides both $|x| = j = n - 2i$ and $|xy| = i$, we have $|s|$ divides $\gcd(i, j) = \gcd(i, n-2i) = \gcd(n, i, n-2i) = \gcd(n, i, j) = g$. Therefore, there exists a length-g word u such that $x = u^{j/g}$ and $y = u^{(i-j)/g}$. We also have $\text{ham}(z, x) = 1$, which implies that there exists a length-g word v with $\text{ham}(u, v) = 1$, and a non-negative integer $p' < j/g$ such that $z = u^{p'}vu^{j/g-p'-1}$. Letting $p = p' + 2i/g$, we have $w = xyxyz = u^{2i/g}u^{p'}vu^{(n-2i)/g-p'-1} = u^pvu^{n/g-p-1}$.

If $xy \neq yx$, then we must have $\text{ham}(xy, yx) = 2$. But since $\text{ham}(xyxyz, yxyzx) = 2$, we must have $\text{ham}(xyz, yzx) = 0$. This means that xyz is a power, but we have already demonstrated that $xyz \in H_2(n-i, i)$. By Corollary 43, this is a contradiction.

$\Longleftarrow$: Let $g = \gcd(n, i, j)$. Suppose we can write $w = u^pvu^{n/g-p-1}$ where $|u| = |v| = g$, and $\text{ham}(u, v) = 1$. Since $g \mid i$, we can write

$$\text{ham}(w, \sigma^i(w)) = \text{ham}(u^pvu^{n/g-p-1}, u^{p-i/g}vu^{n/g+i/g-p-1}) = 2\,\text{ham}(u, v) = 2$$

if $p \leq i/g$, and

$$\text{ham}(w, \sigma^i(w)) = \text{ham}(u^pvu^{n/g-p-1}, u^{n/g-i+p}vu^{p-i-1}) = 2\,\text{ham}(u, v) = 2$$

if $p > i/g$. Since g divides j as well, a similar argument works to show $\text{ham}(w, \sigma^j(w)) = 2$ as well. Therefore, $w \in H_2(n, i)$ and $w \in H_2(n, j)$. $\square$

Lemma 46 shows that any word w that is in $H_2(n,i)$ and $H_2(n,j)$ for $j < i \leq n/2$ is of Hamming distance 1 away from a power. Therefore, to count the number of such words, we need a formula for the number of powers. Recall from Section 2.2, the formulas for $p_k(n)$ and $\psi_k(n)$.

Let $H_2'(n,i)$ denote the set of words $w \in H_2(n,i)$ that are also in $H_2(n,j)$ for some $j < i$. Let $h_2'(n,i) = |H_2'(n,i)|$.

Corollary 47. *Let n, i be positive integers such that $n \geq 2i$. Then*

$$h_2'(n,i) = \begin{cases} n(k-1)p_k(i), & \text{if } i \mid n; \\ n(k-1)k^{\gcd(n,i)}, & \text{otherwise.} \end{cases}$$

Let $H_2''(n,i)$ denote the set of words $w \in H_2(n,i)$ such that $w \notin H_2(n,j)$ for all $j < i$. Let $h_2''(n,i) = |H_2''(n,i)|$.

Lemma 48. *Let n, i be positive integers such that $n > i$. Then*

$$h_2''(n,i) = \begin{cases} \frac{1}{2}n(k-1)\big(k^{\gcd(n,i)}\big(\frac{n}{\gcd(n,i)} - 1\big) - 2p_k(i)\big), & \text{if } i \mid n; \\ \frac{1}{2}k^{\gcd(n,i)}(k-1)n\big(\frac{n}{\gcd(n,i)} - 3\big), & \text{otherwise.} \end{cases}$$

Proof. Let w be a length-n word. The word w is in $H_2''(n,i)$ precisely if it is in $H_2(n,i)$ but not in any $H_2(n,j)$ for $j < i$. So computing $h_2''(n,i)$ reduces to computing the number of length-n words that are in $H_2(n,i)$ and $H_2(n,j)$ for some $j < i$ (i.e., $h_2'(n,i)$) and then subtracting it from the number of words in $H_2(n,i)$ (i.e., $h_2(n,i)$). Therefore

$$h_2''(n,i) = h_2(n,i) - h_2'(n,i) = \begin{cases} \frac{1}{2}n(k-1)\big(k^{\gcd(n,i)}\big(\frac{n}{\gcd(n,i)} - 1\big) - 2p_k(i)\big), & \text{if } i \mid n; \\ \frac{1}{2}k^{\gcd(n,i)}(k-1)n\big(\frac{n}{\gcd(n,i)} - 3\big), & \text{otherwise.} \end{cases}$$

□

Theorem 49. *Let n be an integer ≥ 2. Then*

$$h_2(n) = \sum_{i=1}^{\lfloor n/2 \rfloor} h_2''(n,i).$$

Proof. Every word that is in $H_2(n)$ must also be in $H_2(n,i)$ for some integer i in the range $1 \leq i \leq n-1$. By Lemma 41 we have that every word that is in $H_2(n,i)$ is also in $H_2(n,n-i)$. Therefore we only need to consider words in $H_2(n,i)$ where i is an integer

with $i \leq n-i \implies i \leq n/2$. Consider the quantity $S = \sum_{i=1}^{\lfloor n/2 \rfloor} h_2(n,i)$. Since any member of $H_2(n)$ must also be a member of $H_2(n,i)$ for some $i \leq \lfloor n/2 \rfloor$, we have that $h_2(n) \leq S$. But any member of $H_2(n,i)$ may also be a member of $H_2(n,j)$ for some $j < i$. These words are accounted for multiple times in the sum S. To avoid double-counting we must count the number of words w that are in $H_2(n,i)$ but not in $H_2(n,j)$ for any $j < i$. This quantity is exactly $h_2''(n,i)$. Therefore

$$h_2(n) = \sum_{i=1}^{\lfloor n/2 \rfloor} h_2''(n,i).$$

□

4.7 Exactly one conjugate

So far we have been interested in length-n words u that have at least one conjugate of Hamming distance 2 away from u. But what about length-n words u that have exactly one conjugate of Hamming distance 2 away from u? In this section we provide a formula for the number $h_2'''(n)$ of length-n words u with exactly one conjugate v such that $\text{ham}(u,v) = 2$. See A179677 in the OEIS [95] for the sequence $(h_2'''(n))_{n\geq 0}$.

Let n and i be positive integers such that $n > i$. Let $H_2'''(n)$ denote the set of length-n words u over Σ_k that have exactly one conjugate v with $\text{ham}(u,v) = 2$. Let $h_2'''(n) = |H_2'''(n)|$. Let $H_2''(n,i)$ denote the set of length-n words w such that w is in $H_2(n,i)$ but is not in $H_2(n,j)$ for any $j \neq i$. Let $h_2''(n,i) = |H_2''(n,i)|$.

Suppose $w \in H_2'''(n,i)$. Then by definition we have that $w \in H_2(n,i)$ and $w \notin H_2(n,j)$ for any $j \neq i$. But by Lemma 41 we have that if w is in $H_2(n,i)$ then it must also be in $H_2(n,n-i)$. So if $i \neq n-i$, then w has at least two distinct conjugates of Hamming distance 2 away from it, namely $\sigma^i(w)$ and $\sigma^{n-i}(w)$. Therefore we have $i = n-i$. This implies that n must be even, so $H_2'''(2m+1) = \{\}$ for all $m \geq 1$. Since $i = n-i \implies i = n/2$, we have that $w \in H_2(n,n/2)$. However w cannot be in $H_2(n,j)$ for any $j \neq n/2$. Since any word in $H_2(n,j)$ is also in $H_2(n,n-j)$, the condition of $w \notin H_2(n,j)$ for any $j \neq n/2$ is equivalent to $w \notin H_2(n,j)$ for any j with $1 \leq j < n/2$. But this is just the definition of $H_2''(n,n/2)$. From this we get the following theorem.

Theorem 50. *Let $n \geq 1$ be an integer. Then*

$$h_2'''(n) = \begin{cases} \frac{1}{2}n(k-1)(k^{n/2} - 2p_k(n/2)), & \text{if } n \text{ is even;} \\ 0, & \text{otherwise.} \end{cases}$$

4.8 Lyndon conjugates

A *Lyndon word* is a word that is lexicographically smaller than any of its non-trivial conjugates. In this section we count the number of Lyndon words in $H_2(n)$. See A226893 in the OEIS [95] for this sequence.

Theorem 51. *There are* $\frac{h_2(n)}{n}$ *Lyndon words in* $H_2(n)$.

Proof. Corollary 44 says that all members of $H_2(n)$ are primitive and Lemma 45 says that if a word is in $H_2(n)$, then any conjugate of it is also in $H(n)$. It is easy to verify that every primitive word has exactly one Lyndon conjugate. Therefore exactly $\frac{h_2(n)}{n}$ words in $H_2(n)$ are Lyndon words. □

4.9 Asymptotic behaviour of almost-commuting words

In this section we show that $h_2(n)$ grows erratically. We do this by demonstrating that $h(n)$ is a cubic polynomial for prime n, and that $h_2(n)$ is bounded below by an exponential for even n.

Lemma 52. *Let* n *be a prime number. Then*

$$h_2(n) = \frac{1}{4}k(k-1)n(n^2-4n+7).$$

Proof. Let $n > 1$ be a prime number. Since n is prime, we have that $\gcd(n,i) = 1$ for all integers i with $1 < i < n$. Then

$$\begin{aligned}
h_2(n) &= \sum_{i=1}^{(n-1)/2} h_2''(n,i) \\
&= \frac{1}{2}k(k-1)n(n-1) + \sum_{i=2}^{(n-1)/2} \frac{1}{2}k^{\gcd(n,i)}(k-1)n\left(\frac{n}{\gcd(n,i)} - 3\right) \\
&= \frac{1}{2}k(k-1)n(n-1) + \left(\frac{n-3}{2}\right)\frac{1}{2}k(k-1)n(n-3) \\
&= \frac{1}{4}k(k-1)n(n^2-4n+7).
\end{aligned}$$

□

Lemma 53. *Let $n > 1$ be an integer. Then $h_2(2n) \geq nk^n$.*

Proof. Since any word in $H_2(2n, n)$ must also be in $H_2(2n)$, we have that $h_2(2n) \geq h_2(2n, n)$. From Lemma 36 we see that

$$h_2(2n, n) = \frac{1}{2}k^{\gcd(2n,n)}(k-1)2n\left(\frac{2n}{\gcd(2n, n)} - 1\right) = k^n(k-1)n.$$

Since $k \geq 2$, we have that $k - 1 \geq 1$. Therefore $h_2(2n) \geq k^n(k-1)n \geq nk^n$ for all $n > 1$. □

Chapter 5

Words with exactly one border

5.1 Introduction

Earlier in Section 2.3, words with a unique border were mentioned as words where the shortest period and the length of the shortest border are equal. Harju and Nowotka [58] counted the number $B_k(n)$ of length-n words with a unique border, and $B_k(n,t)$ of length-n words over a k-letter alphabet with a fixed length-t unique border; see Theorem 54.

Theorem 54 (Harju and Nowotka [58])**.** *Let n, t, and k be integers such that $k \geq 2$ and $n \geq 2t \geq 2$. Then*

$$B_k(n,t) = u_t(k^{n-2t} - W_k(n-2t,t) - E_k(n-2t,t))$$

where

$$W_k(r,s) = \begin{cases} k^{r-2s} - u_{r-2s}, & \text{if } r > 2s; \\ 1, & \text{if } r = 2s; \\ 0, & \text{otherwise.} \end{cases}$$

and

$$E_k(r,s) = \begin{cases} k^{(r-s)/2}, & \text{if } s < r < 3s \text{ and } r-s \text{ is even;} \\ 1, & \text{if } r = s; \\ 0, & \text{otherwise.} \end{cases}$$

Through personal communication with the authors, a small error in one of the proofs leading up to Theorem 54 was discovered. In this chapter we present the correct recurrence

for the number of length-n words with a unique border. We also show that the probability a length-n word has a unique border tends to a constant. See A334600 in the *On-Line Encyclopedia of Integer Sequences* (OEIS) [95] for the sequence $(B_2(n))_{n\geq 0}$.

5.2 Counting words with unique borders

To solve the first problem, we use similar ideas to the ideas used to enumerate bordered words and mutually bordered pairs of words. Since the shortest border of a word is unbordered, one has that the unique border of a word must be unbordered. It is also true that the shortest border of a word cannot exceed half the length of the word. By combining these ideas we get Theorem 55 and Theorem 56.

Theorem 55. *Let $n > t \geq 1$ be integers. Then the number of length-n words with a unique length-t border is*

$$B_k(n,t) = \begin{cases} 0, & \text{if } n < 2t; \\ u_t k^{n-2t} - \sum_{i=2t}^{\lfloor n/2 \rfloor} B_k(i,t) k^{n-2i}, & \text{if } n \geq 2t \text{ and } n+t \text{ odd;} \\ u_t k^{n-2t} - B_k((n+t)/2,t) - \sum_{i=2t}^{\lfloor n/2 \rfloor} B_k(i,t) k^{n-2i}, & \text{if } n \geq 2t \text{ and } n+t \text{ even.} \end{cases}$$

Proof. Let w be a length-n word with a unique length-t border u. Since u is the unique border of w, it is unbordered. Thus we can write $w = uvu$ for some (possibly empty) word v. If $n < 2t$ then $B_k(n,t) = 0$ since u is unbordered and thus cannot overlap itself in w.

Suppose $n \geq 2t$. Let $\overline{B_k}(n,t)$ denote the number of length-n words that have a length-t unbordered border and have a border of length $> t$. Clearly $B_k(n,t) = u_t k^{n-2t} - \overline{B_k}(n,t)$. Suppose w as another border u' of length $> t$. Furthermore suppose that there is no other border u'' with $|u| < |u''| < |u'|$. Then u is the largest border of u'. Since u is the shortest border, we have $|u| \leq n/2$. But we could possibly have $|u'| > n/2$. The only possible way for $|u'|$ to exceed $n/2$ is if $w = uv'uv'u$ for some (possibly empty) word v. But this is only possible if $n+t$ is even, otherwise we cannot place u in the centre of w. When $n+t$ is odd, we can compute $\overline{B_k}(n,t)$ by summing over all possibilities for u' (i.e., $2t \leq |u'| \leq \lfloor n/2 \rfloor$) and the middle part of w (i.e., v'' where $w = u'v''u'$). This translates to

$$\overline{B_k}(n,t) = \sum_{i=2t}^{\lfloor n/2 \rfloor} B_k(i,t) k^{n-2i}.$$

When $n+t$ is even, we compute $\overline{B_k}(n,t)$ the same, except we also include the case where $|u'| = (n+t)/2$. This translates to

$$\overline{B_k}(n,t) = B_k((n+t)/2, t) + \sum_{i=2t}^{\lfloor n/2 \rfloor} B_k(i,t)k^{n-2i}.$$

□

Theorem 56. *Let $n \geq 2$ be an integer. Then the number of length-n words with a unique border is*

$$B_k(n) = \sum_{t=1}^{\lfloor n/2 \rfloor} B_k(n,t).$$

5.3 Limiting values

In this section we show that the probability that a random word of length n has a unique border tends to a constant. Table 5.1 shows the behaviour of this probability as k increases.

The probability that a random word of length n has a unique border corresponds to the sum

$$P_{n,k} = \frac{B_k(n)}{k^n} = \frac{1}{k^n} \sum_{i=1}^{\lfloor n/2 \rfloor} B_k(n,i).$$

Lemma 57. *Let $k \geq 2$ and $n \geq 2t \geq 2$ be integers. Then*

$$\frac{B_k(n,t)}{k^n} \leq \frac{1}{k^t}.$$

Proof. Let w be a length-n word. Suppose w has a unique border of length t. Since $t \leq n/2$, we can write $w = uvu$ for some words u and v where $|u| = t$. But this means that $B_k(n,t) \leq k^{n-t}$, and the lemma follows. □

Theorem 58. *Let $k \geq 2$ be an integer. Then the limit $P_k = \lim_{n\to\infty} P_{n,k}$ exists.*

Proof. Follows from the definition of $P_{n,k}$, Lemma 57, and the direct comparison test for convergence. □

k	$\approx P_k$
2	0.5155
3	0.3910
4	0.2922
5	0.2302
6	0.1890
7	0.1599
8	0.1384
9	0.1219
10	0.1089
$\vdots$	$\vdots$
100	0.0101

Table 5.1: Probability that a word has a unique border.

Chapter 6

Block palindromes

6.1 Introduction

A *palindrome* is a word that reads the same forwards as it does backwards. More formally, letting $w^R = w_n w_{n-1} \cdots w_1$ where $w = w_1 w_2 \cdots w_n$, a palindrome is a word w such that $w = w^R$. The definition of a palindrome is quite restrictive. The second half of a palindrome is fully determined by the first half. Thus, compared to all length-n words, the number of length-n palindromes is vanishingly small. But many words exhibit palindrome-like structure. Take the English word `marjoram`. It is clearly not a palindrome but it comes close. Replacing the block `jo` with a single letter turns the word into a palindrome. In this chapter we consider a generalization of palindromes that incorporates this kind of palindromic structure.

In the 2015 British Olympiad [1], the concept of a *block palindrome* was first introduced. Let w be a non-empty word. A block palindrome of w is a factorization $w = w_{-m} \cdots w_{-1} w_0 w_1 \cdots w_m$ of a word such that w_0 is a possibly empty word, and every other factor $w_{-i} = w_i$ is non-empty for all i with $1 \leq i \leq m$. We say that a block palindrome $w_{-m} \cdots w_{-1} w_0 w_1 \cdots w_m$ is of *width* t where $t = 2m + 1$ if w_0 is non-empty and $t = 2m$ otherwise. In other words, the width of a block palindrome is the number of blocks in the factorization. The *largest block palindrome*[1] [49] of a word w is a block palindrome $w = w_{-m} \cdots w_{-1} w_0 w_1 \cdots w_m$ where m is maximized (i.e., where the width of the block palindrome is maximized). See [74] for more on the topic of block palindromes

[1]Largest block palindromes also appear in https://www.reddit.com/r/math/comments/ga2iyo/i_just_defined_the_palindromity_function_on/.

and block reversal. Kolpakov and Kucherov [70] studied a special case of block palindromes, the *gapped palindrome*. If w_0 is non-empty and $|w_{-i}| = |w_i| = 1$ for all i with $1 \leq i \leq m$, then w is said to be a *gapped palindrome*. Régnier [84] studied something similar to the block palindromes, but in her paper she was concerned with borders of borders, and not iteratively "peeling" off borders. See [41, 83] for results on factoring words into palindromes.

Example 59. We use the centre dot $\cdot$ to denote the separation between blocks in the block palindrome of a word.

Consider the word `abracadabra`. It has the following block palindromes:

$$\texttt{abracadabra},$$
$$\texttt{abra} \cdot \texttt{cad} \cdot \texttt{abra},$$
$$\texttt{a} \cdot \texttt{br} \cdot \texttt{a} \cdot \texttt{cad} \cdot \texttt{a} \cdot \texttt{br} \cdot \texttt{a}.$$

The last block palindrome is of width 7 and has the longest width; thus it is the largest block palindrome of `abracadabra`.

Consider the word `reappear`. It has the following block palindromes:

$$\texttt{reappear},$$
$$\texttt{r} \cdot \texttt{eappea} \cdot \texttt{r},$$
$$\texttt{r} \cdot \texttt{ea} \cdot \texttt{pp} \cdot \texttt{ea} \cdot \texttt{r},$$
$$\texttt{r} \cdot \texttt{ea} \cdot \texttt{p} \cdot \texttt{p} \cdot \texttt{ea} \cdot \texttt{r}.$$

The last block palindrome is of width 6 and has the longest width; thus it is the largest block palindrome of `reappear`.

Let w be a length-n word. Suppose $w_{-m} \cdots w_{-1} w_0 w_1 \cdots w_m$ is the largest block palindrome of w. Goto et al. [49] showed that w_i is the shortest border of $w_{-i} \cdots w_{-1} w_0 w_1 \cdots w_i$. This means that we can compute the the largest blocked palindrome of w by greedily "peeling off" the shortest borders of central factors.

Let $\mathrm{LBP}_k(n, t)$ denote the number of length-n words with a width-t largest block palindrome.

The rest of the chapter is structured as follows. In Section 6.2 we give a recurrence for $\mathrm{LBP}_k(n, t)$; see Theorem 60. In Section 6.3 we show that the expected width of the largest block palindrome of a length-n word tends to a constant; see Theorem 62. Finally, in Section 6.4 we consider *smallest* block palindromes in the sense that one "peels off" the largest non-overlapping border.

6.2 Counting largest block palindromes

In this section we prove a recurrence for $\mathrm{LBP}_k(n,t)$. See Table 6.1 for sample values of $\mathrm{LBP}_2(n,t)$ for small n, t. Recall that u_n denotes the number of length-n unbordered words over a k-letter alphabet. See Section 2.4 for a recurrence for u_n.

Theorem 60. *Let $n, t \geq 0$, and $k \geq 2$ be integers. Then*

$$\mathrm{LBP}_k(n,t) = \begin{cases} \sum_{i=1}^{(n-t)/2+1} u_i \,\mathrm{LBP}_k(n-2i, t-2), & \text{if } n,\ t \text{ even;} \\ \sum_{i=1}^{(n-t+1)/2} u_{2i} \,\mathrm{LBP}_k(n-2i, t-1), & \text{if } n \text{ even},\ t \text{ odd;} \\ 0, & \text{if } n \text{ odd},\ t \text{ even;} \\ \sum_{i=1}^{(n-t)/2+1} u_{2i-1} \,\mathrm{LBP}_k(n-2i+1, t-1), & \text{if } n,\ t \text{ odd.} \end{cases}$$

where

$$\begin{aligned} \mathrm{LBP}_k(0,0) &= 0, \\ \mathrm{LBP}_k(2n,2) &= u_n, \\ \mathrm{LBP}_k(n,1) &= u_n. \end{aligned}$$

Proof. Let w be a length-n word. Suppose w has a width-t largest block palindrome $w_{-m} \cdots w_{-1} w_0 w_1 \cdots w_m$ where $t = 2m+1$ iff w_0 is non-empty and $t = 2m$ otherwise. Clearly $\mathrm{LBP}_k(0,0) = 0$. We know that each block in a largest block palindrome is unbordered, since it is a shortest border. This immediately implies $\mathrm{LBP}_k(n,1) = u_n$ and $\mathrm{LBP}_k(2n,2) = u_n$.

Now we take care of the other cases.

- Suppose n, t are even. Then by removing w_1 and w_{-1} from w, we get $w' = w_{-m} \cdots w_{-2} w_2 \cdots w_m$, which is a length-$(n - 2|w_1|)$ word with a largest block palindrome of width $t-2$. This mapping is clearly reversible since all blocks in a largest block palindrome are unbordered, including w_1. Thus summing over all possible w_1 and all length-$(n-2|w_1|)$ words with a largest block palindrome of width $t-2$ we have
$$\mathrm{LBP}_k(n,t) = \sum_{i=1}^{(n-t)/2+1} u_i \,\mathrm{LBP}_k(n-2i, t-2).$$
- Suppose n is even and t is odd. Then by removing w_0 from w, we get $w' = w_{-m} \cdots w_{-1} w_1 \cdots w_m$, which is a length-$(n - |w_0|)$ word with a largest block palindrome of width $t-1$. This mapping is reversible for the same reason as in the previous

case. The word w' is of even length since $|w'| = 2|w_1 \cdots w_m|$. Since n is even and $|w'|$ is even, we must have that $|w_0|$ is even as well. Thus summing over all possible w_0 and all length-$(n - |w_0|)$ words with a largest block palindrome of width $t-1$ we have

$$\mathrm{LBP}_k(n,t) = \sum_{i=1}^{(n-t+1)/2} u_{2i}\,\mathrm{LBP}_k(n-2i, t-1).$$

- Suppose n is odd and t is even. Then the length of w is $2|w_1 \cdots w_m|$, which is even, a contradiction. Thus $\mathrm{LBP}_k(n,t) = 0$.

- Suppose n, t are odd. Then by removing w_0 from w, we get $w' = w_{-m} \cdots w_{-1} w_1 \cdots w_m$, which is a length-$(n-|w_0|)$ word with a largest block palindrome of width $t-1$. This mapping is reversible for the same reasons as in the previous cases. Since n is odd and $|w'|$ is even (proved in the previous case), we must have that $|w_0|$ is odd. Thus summing over all possible w_0 and all length-$(n - |w_0|)$ words with a largest block palindrome of width $t-1$ we have

$$\sum_{i=1}^{(n-t)/2+1} u_{2i-1}\,\mathrm{LBP}_k(n-2i+1, t-1).$$

□

n \ t	1	2	3	4	5	6	7	8	9	10
10	284	12	224	40	168	72	96	64	32	32
11	568	0	472	0	416	0	336	0	192	0
12	1116	20	856	88	656	176	448	224	224	160
13	2232	0	1752	0	1488	0	1248	0	896	0
14	4424	40	3328	176	2544	432	1856	640	1152	640
15	8848	0	6736	0	5440	0	4576	0	3584	0
16	17622	74	13100	372	9896	984	7408	1744	5088	2080
17	35244	0	26348	0	20536	0	16784	0	13664	0
18	70340	148	51936	760	38824	2248	29152	4416	21088	6240
19	140680	0	104168	0	79168	0	62800	0	51008	0
20	281076	284	206744	1592	153344	4992	114688	10912	84704	17312

Table 6.1: Some values of $\mathrm{LBP}_2(n,t)$ for n, t where $10 \le n \le 20$ and $1 \le t \le 10$.

6.3 Expected width of largest block palindrome

In this section we show that the expected width of the largest block palindrome of a length-n word is bounded by a constant. Table 6.2 shows the behaviour of this expected value as k increases.

The expected width of the largest block palindrome of a length-n word corresponds to

$$E_{n,k} = \frac{1}{k^n} \sum_{i=1}^{n} i \cdot \mathrm{LBP}_k(n, i).$$

Lemma 61. *Let $k \geq 2$ and $n \geq t \geq 1$ be integers. Then*

$$\frac{\mathrm{LBP}_k(n,t)}{k^n} \leq \frac{1}{k^{t/2-1}}.$$

Proof. Let w be a length-n word. Suppose w has a width-t largest block palindrome $w_{-m} \cdots w_{-1} w_0 w_1 \cdots w_m$ where $t = 2m+1$ iff w_0 is non-empty and $t = 2m$ otherwise. The blocks $w_1, w_2, \ldots, w_m$ are all determined by $w_{-1}, w_{-2}, \ldots, w_{-m}$. Since w_i is non-empty for every $1 \leq i \leq m$, we have that $\mathrm{LBP}_k(n,t) \leq k^{n-m} \leq k^{n-t/2+1}$. So

$$\frac{\mathrm{LBP}_k(n,t)}{k^n} \leq \frac{1}{k^{t/2-1}}$$

for all $n \geq t \geq 1$. □

Theorem 62. *The limit $E_k = \lim_{n\to\infty} E_{n,k}$ exists for all $k \geq 2$.*

Proof. Follows from the definition of $E_{n,k}$, Lemma 61, and the direct comparison test for convergence. □

k	$\approx E_k$
2	6.4686
3	2.5908
4	1.9080
5	1.6314
6	1.4827
7	1.3902
8	1.3272
9	1.2817
10	1.2472
⋮	⋮
100	1.0204

Table 6.2: Asymptotic expected length of a word's largest block palindrome.

In [28], the authors prove that the expected length of the longest unbordered factor in a word is $\Theta(n)$. Taking this into account, it is not surprising that the expected length of the largest block palindrome of a word tends to a constant.

6.4 Smallest block palindrome

A word w, seen as a block, clearly satisfies the definition of a block palindrome. Thus, taken literally, the smallest block palindrome for all words is of width 1. But this is not very interesting, so we adjust the definition of the smallest block palindrome. We say that the *smallest block palindrome* of a word w is a block palindrome $w = w_{-m} \cdots w_{-1} w_0 w_1 \cdots w_m$ where each w_i is the largest non-overlapping border of $w_{-i} \cdots w_{-1} w_0 w_1 \cdots w_m$, except w_0, which is either empty or unbordered. For example, going back to the example in the introduction, the smallest block palindrome of `abracadabra` is $\texttt{abra} \cdot \texttt{cad} \cdot \texttt{abra}$ and the smallest block palindrome of `reappear` is $\texttt{r} \cdot \texttt{ea} \cdot \texttt{p} \cdot \texttt{p} \cdot \texttt{ea} \cdot \texttt{r}$.

A natural question ask is: what is the maximum width $f_k(n)$ of the smallest block palindrome of a word over Σ_k?

- Jeffrey Shallit conjectured that $f_2(8n + i) = 6n + i$ for i with $0 \leq i \leq 5$ and $f_2(8n + 6) = f_2(8n + 7) = 6n + 5$. He also conjectured that $f_k(n) = n$ for $k \geq 3$.

To calculate $f_k(n)$, two things are needed. One needs an upper bound for $f_k(n)$. One also requires words that witness the upper bound. Jeffrey Shallit found the words that witness the upper bound in the following theorems.

Theorem 63. *Let $n \geq 0$ be an integer. Then $f_2(8n+i) = 6n+i$ for i with $0 \leq i \leq 5$ and $f_2(8n+6) = f_2(8n+7) = 6n+5$.*

Proof. We start by proving lower bounds on $f_2(n)$. Let $m \geq 0$ be an integer. Suppose $m = 8n$. Then the width of the smallest block palindrome of

$$(0101)^n(1001)^n$$

is $6n$. To see this, notice that the smallest block palindrome of 01011001 is of width 6. Suppose $m = 8n+i$ for some i with $1 \leq i \leq 7$. Then one can take $(0101)^n(1001)^n$ and insert either 0, 00, 010, 0110, 01010, 010110, or 0110110 to the middle of the word to get the desired length.

Now we prove upper bounds on $f_2(n)$. Let $t \leq n$ be a positive integer. Let w be a length-n word. Suppose w has a width-t largest block palindrome $w_{-m} \cdots w_{-1} w_0 w_1 \cdots w_m$ where $t = 2m+1$ iff w_0 is non-empty and $t = 2m$ otherwise. One can readily verify $f_k(n)$ through exhaustive search of all binary words of length < 8. Suppose $m \geq 4$, so $n \geq t \geq 8$. Then we can write $w = w_{-m} w_{-m+1} w_{-m+2} \cdots w_{m-2} w_{m-1} w_m$ where $|w_{m-2}|, |w_{m-1}|, |w_m| > 0$. It is easy to show that $|w_{m-2} w_{m-1} w_m| \geq 4$ by checking that all binary words of length < 8 do not admit a smallest block palindrome of width 6. The upper bound immediately follows from this. In the worst case we can peel off prefixes and suffixes of length 4 while accounting for the 6 blocks they add to the block palindrome until we hit the middle core of length < 8. □

Theorem 64. *Let $n \geq 0$ and $k \geq 3$ be integers. Then $f_k(n) = n$.*

Proof. Clearly $f_k(n) \leq n$. We prove $f_k(n) \geq n$. If n is divisible by 6, then consider the word $(012)^{n/6}(210)^{n/6}$. If n is not divisible by 6, then take $(012)^{\lfloor n/6 \rfloor}(210)^{\lfloor n/6 \rfloor}$ and insert either 0, 00, 010, 0110, or 01010 in the middle of the word, When calculating the smallest block palindrome of the resulting words, it is easy to see that at each step we are removing a border of length 1. Thus their largest block palindrome is of width n. □

Chapter 7

Bounds for the number of closed and privileged words

7.1 Introduction

In this chapter we present two of the main results of this thesis. We present improved bounds on the number of closed and privileged words. The main two results of this chapter are Theorem 65 and Theorem 66.

Recall that a word w is said to be *closed* if $|w| \leq 1$ or if w has a border that occurs exactly twice in w. If u is a border w and u occurs in w exactly twice, then we say w is *closed by* u. It is easy to see that if a word w is closed by a word u, then u must be the largest border in w. Otherwise u would occur more than two times in w. A word w is said to be *privileged* if $|w| \leq 1$ or if w is closed by a privileged word. See Example 17 for examples illustrating these definitions.

Both closed words [37] and privileged words [67] have been introduced relatively recently, although some equivalent formulations of closed words that have been defined previously; see Section 2.4 for more information on these equivalent formulations.

Since their introduction, there has been much research into the properties of closed and privileged words [81, 24, 31, 6, 89, 38, 66]. One problem that has received some interest lately [40, 79, 86, 87] is to find good upper and lower bounds for the number of closed and privileged words.

Let $C_k(n)$ denote the number of length-n closed words over Σ_k. Let $C_k(n, t)$ denote the number of length-n closed words over Σ_k that are closed by a length-t word. Let $P_k(n)$

denote the number of length-n privileged words over Σ_k. Let $P_k(n,t)$ denote the number of length-n privileged words over Σ_k that are closed by a length-t privileged word. See Tables 7.1, 7.2, and 7.3 for sample values of $C_2(n)$, $C_2(n,t)$, $P_2(n)$, and $P_2(n,t)$ for small n, t. See sequences A226452 and A231208 in the *On-Line Encyclopedia of Integer Sequences* (OEIS) [95].

Every privileged word is a closed word, so any upper bound on $C_k(n)$ is also an upper bound on $P_k(n)$. Furthermore, any lower bound on $P_k(n)$ is also a lower bound on $C_k(n)$. See Section 2.4 for a brief literature review on the best known bounds on $C_k(n)$ and $P_k(n)$.

The best upper and lower bounds for both $C_k(n)$ and $P_k(n)$ are widely separated, and can be much improved. In this chapter we completely resolve the asymptotic behaviour of the number of length-n closed words, by showing that it is asymptotically $\Theta(\frac{k^n}{n})$. Additionally, we nearly completely resolve the asymptotic behaviour of the number of length-n privileged words, by giving a family of upper and lower bounds that are separated by a factor that grows arbitrarily slowly. We prove the following two theorems.

Theorem 65. *Let $k \geq 2$ be an integer.*

(a) There exist constants N and c such that $C_k(n) \geq c\frac{k^n}{n}$ for all $n > N$.

(b) There exist constants N' and c' such that $C_k(n) \leq c'\frac{k^n}{n}$ for all $n > N'$.

Theorem 66. *Let $k \geq 2$ be an integer. Let $\log_k^{\circ 0}(n) = n$ and $\log_k^{\circ j}(n) = \log_k(\log_k^{\circ j-1}(n))$ for $j \geq 1$.*

(a) For all $j \geq 0$ there exist constants N_j and c_j such that

$$P_k(n) \geq c_j \frac{k^n}{n \log_k^{\circ j}(n) \prod_{i=1}^{j} \log_k^{\circ i}(n)}$$

for all $n > N_j$.

(b) For all $j \geq 0$ there exist constants N'_j and c'_j such that

$$P_k(n) \leq c'_j \frac{k^n}{n \prod_{i=1}^{j} \log_k^{\circ i}(n)}$$

for all $n > N'_j$.

Before we proceed, we give a heuristic argument as to why $C_k(n)$ is in $\Theta(\frac{k^n}{n})$. Consider a "random" length-n word w. Let $\ell = \log_k(n) + c$. The probability that w has a length-ℓ border u is around $k^{n-\ell}/k^n = \frac{1}{k^c n}$. Suppose w has a length-ℓ border. Now suppose we drop the first and last character of w to get w'. If w' were randomly chosen (which it is not), then we could use the linearity of expectation to get that the expected number of occurrences of u in w' is approximately $(n-2)k^{-\ell} \approx k^{-c}$. Thus for c large enough we have that u does not occur in w with high probability, and so w is closed. Therefore there are approximately $k^{n-\ell} \in \Theta(\frac{k^n}{n})$ length-n closed words.

n \ t	1	2	3	4	5	6	7	8	9	10
10	2	30	70	50	30	12	6	2	2	0
11	2	42	118	96	54	30	13	6	2	2
12	2	60	200	182	114	54	30	12	6	2
13	2	88	338	346	214	126	54	30	12	6
14	2	132	570	640	432	232	126	54	30	12
15	2	202	962	1192	828	474	240	126	54	30
16	2	314	1626	2220	1612	908	492	240	126	54
17	2	494	2754	4128	3112	1822	956	504	240	126
18	2	784	4676	7670	6024	3596	1934	982	504	240
19	2	1252	7960	14264	11636	7084	3828	1992	990	504
20	2	2008	13588	26524	22512	13928	7632	3946	2026	990

Table 7.1: Some values of $C_2(n, t)$ for n, t where $10 \leq n \leq 20$ and $1 \leq t \leq 10$.

n \ t	1	2	3	4	5	6	7	8	9	10
10	2	16	22	8	6	2	2	0	2	0
11	2	26	38	16	10	6	4	2	2	2
12	2	42	68	30	18	4	6	2	2	0
13	2	68	122	58	38	14	10	6	4	2
14	2	110	218	108	76	20	14	8	6	2
15	2	178	390	204	148	46	24	18	14	6
16	2	288	698	384	288	86	48	16	18	8
17	2	466	1250	724	556	178	92	36	32	26
18	2	754	2240	1364	1076	344	190	64	36	28
19	2	1220	4016	2572	2092	688	388	136	70	56
20	2	1974	7204	4850	4068	1342	772	268	138	52

Table 7.2: Some values of $P_2(n, t)$ for n, t where $10 \leq n \leq 20$ and $1 \leq t \leq 10$.

n	$P_2(n)$	$C_2(n)$	n	$P_2(n)$	$C_2(n)$
0	1	1	13	328	1220
1	2	2	14	568	2240
2	2	2	15	1040	4132
3	4	4	16	1848	7646
4	4	6	17	3388	14244
5	8	12	18	6132	26644
6	8	20	19	11332	49984
7	16	36	20	20788	94132
8	20	62	21	38576	177788
9	40	116	22	71444	336756
10	60	204	23	133256	639720
11	108	364	24	248676	1218228
12	176	664	25	466264	2325048

Table 7.3: Some values of $P_2(n)$ and $C_2(n)$ for $n \leq 25$.

7.2 Preliminary results

In this section we give some necessary results and definitions in order to prove our main results. Also throughout this chapter, we use c's, d's, and N's to denote positive real constants (dependent on k).

Let w be a length-n word. Suppose w is closed by a length-t word u. Since u is also the largest border of w, it follows that w cannot be closed by another word. This implies that

$$C_k(n) = \sum_{i=1}^{n-1} C_k(n,t) \text{ and } P_k(n) = \sum_{i=1}^{n-1} P_k(n,t)$$

for $n > 1$.

Let $B_k(n,u)$ denote the number of length-n words over Σ_k that are closed by the word u. Let $A_k(n,u)$ denote the number of length-n words over Σ_k that do not contain the word u as a factor.

Recall from Section 2.4 that the *auto-correlation* of a length-t word u is a length-t binary word $a(u) = a_1 a_2 \cdots a_t$ where $a_i = 1$ if and only if u has a border of length $t-i+1$. The *auto-correlation polynomial* [50, 51, 52] $f_{a(u)}(z)$ of $a(u)$ is defined as

$$f_{a(u)}(z) = \sum_{i=0}^{t-1} a_{t-i} z^i.$$

For example, the word $u = \texttt{alfalfa}$ has auto-correlation $a(u) = 1001001$ and auto-correlation polynomial $f_{a(u)}(z) = z^6 + z^3 + 1$.

We now prove two technical lemmas that will be used in the proofs of Theorem 65 (b) and Theorem 66 (b).

Lemma 67. *Let $k, t \geq 2$ be integers, and let γ be a real number such that $0 < \gamma \leq \frac{6}{t}$. Then*

$$k^t - \gamma t k^{t-1} \leq (k-\gamma)^t \leq k^t - \gamma t k^{t-1} + \frac{1}{2}\gamma^2 t(t-1)k^{t-2}.$$

Proof. The case when $k = 2$ was proved in a paper by Forsyth et al. [40, Lemma 9]. We generalize their proof to $k \geq 3$.

When $t = 2$, we have $k^2 - 2k\gamma \le (k-\gamma)^2 \le k^2 - 2k\gamma + \gamma^2$. So suppose $t \ge 3$. By the binomial theorem, we have

$$\begin{aligned}(k-\gamma)^t &= \sum_{i=0}^{t} k^{t-i}(-\gamma)^i \binom{t}{i} = k^t - \gamma t k^{t-1} + \sum_{i=2}^{t} k^{t-i}(-\gamma)^i \binom{t}{i} \\ &\ge k^t - \gamma t k^{t-1} + \sum_{j=1}^{\lfloor (t-1)/2 \rfloor} \left(k^{t-2j}\gamma^{2j}\binom{t}{2j} - k^{t-2j-1}\gamma^{2j+1}\binom{t}{2j+1} \right).\end{aligned}$$

So to show that $k^t - \gamma t k^{t-1} \le (k-\gamma)^t$, it is sufficient to show that

$$k^{t-2j}\gamma^{2j}\binom{t}{2j} \ge k^{t-2j-1}\gamma^{2j+1}\binom{t}{2j+1} \tag{7.1}$$

for $1 \le j \le \lfloor (t-1)/2 \rfloor \le (t-1)/2$.

By assumption we have that $\gamma \le \frac{6}{t}$, so $\gamma \le \frac{6}{t-2}$ and thus $\gamma t - 2\gamma \le 6$. Adding $2\gamma - 2$ to both sides we get $\gamma t - 2 \le 4 + 2\gamma$, and so $\frac{\gamma t - 2}{\gamma + 2} \le 2$. If $i \ge 2 \ge \frac{\gamma t-2}{\gamma+2}$, then $(\gamma+2)i \ge \gamma t - 2$. This implies that $2(i+1) \ge \gamma(t-i)$, and

$$\frac{k}{\gamma} \ge \frac{2}{\gamma} \ge \frac{t-i}{i+1} = \frac{\binom{t}{i+1}}{\binom{t}{i}}.$$

Therefore letting $i = 2j$, we have that $k\binom{t}{2j} \ge \gamma\binom{t}{2j+1}$. Multiplying both sides by $k^{t-2j-1}\gamma^{2j}$ we get $k^{t-2j}\gamma^{2j}\binom{t}{2j} \ge k^{t-2j-1}\gamma^{2j+1}\binom{t}{2j+1}$, which proves (7.1).

Now we prove that $(k-\gamma)^t \le k^t - \gamma t k^{t-1} + \frac{1}{2}\gamma^2 t(t-1)k^{t-2}$. Going back to the binomial expansion of $(k-\gamma)^t$, we have

$$\begin{aligned}(k-\gamma)^t &= k^t - \gamma t k^{t-1} + \frac{1}{2}\gamma^2 t(t-1)k^{t-2} + \sum_{i=3}^{t} k^{t-i}(-\gamma)^i \binom{t}{i} \\ &\le k^t - \gamma t k^{t-1} + \frac{1}{2}\gamma^2 t(t-1)k^{t-2} \\ &\quad - \sum_{j=1}^{\lfloor (t-2)/2 \rfloor} \left(k^{t-2j-1}\gamma^{2j+1}\binom{t}{2j+1} - k^{t-2j-2}\gamma^{2j+2}\binom{t}{2j+2} \right).\end{aligned}$$

So to show that $(k-\gamma)^t \le k^t - \gamma t k^{t-1} + \frac{1}{2}\gamma^2 t(t-1)k^{t-2}$, it is sufficient to show that

$$k^{t-2j-1}\gamma^{2j+1}\binom{t}{2j+1} \ge k^{t-2j-2}\gamma^{2j+2}\binom{t}{2j+1}$$

for $1 \le j \le \lfloor (t-2)/2 \rfloor$. But we already have already proved that $k\binom{t}{i} \ge \gamma\binom{t}{i+1}$. Letting $i = 2j$, we have that $k\binom{t}{2j+1} \ge \gamma\binom{t}{2j+2}$. Multiplying both sides by $k^{t-2j-2}\gamma^{2j+1}$ we get $k^{t-2j-1}\gamma^{2j+1}\binom{t}{2j+1} \ge k^{t-2j-2}\gamma^{2j+2}\binom{t}{2j+2}$.

□

Let $\log_k^{\circ 0}(n) = n$ and $\log_k^{\circ j}(n) = \log_k(\log_k^{\circ j-1}(n))$ for $j \ge 1$.

Lemma 68. *Let $i \ge 1$ and $k \ge 2$ be integers. Then for any constant $\gamma > 0$, we have*

$$\lim_{n\to\infty} \frac{\log_k^{\circ i}(n^\gamma)}{\log_k^{\circ i}(n)} = \begin{cases} \gamma, & \text{if } i = 1; \\ 1, & \text{if } i > 1. \end{cases}$$

Proof. When $i = 1$ we have $\lim_{n\to\infty} \frac{\log_k(n^\gamma)}{\log_k(n)} = \gamma \lim_{n\to\infty} \frac{\log_k(n)}{\log_k(n)} = \gamma$.

The proof is by induction on i. Since we will use L'Hôpital's rule to evaluate the limit, we first compute the derivative of $\log_k^{\circ i}(n^\lambda)$ with respect to n for any constant $\lambda > 0$. We have

$$\frac{d}{dn} \log_k^{\circ i}(n^\lambda) = \frac{\lambda}{n \prod_{j=1}^{i-1} \log_k^{\circ j}(n^\lambda)}.$$

In the base case, when $i = 2$, we have

$$\lim_{n\to\infty} \frac{\log_k^{\circ 2}(n^\gamma)}{\log_k^{\circ 2}(n)} = \lim_{n\to\infty} \frac{\frac{\gamma}{n\log_k(n^\gamma)}}{\frac{1}{n\log_k(n)}} = 1.$$

Suppose $i > 2$. Then we have

$$\lim_{n\to\infty} \frac{\log_k^{\circ i}(n^\gamma)}{\log_k^{\circ i}(n)} = \lim_{n\to\infty} \frac{\frac{\gamma}{n \prod_{j=1}^{i-1} \log_k^{\circ j}(n^\gamma)}}{\frac{1}{n \prod_{j=1}^{i-1} \log_k^{\circ j}(n)}} = \lim_{n\to\infty} \frac{\prod_{j=2}^{i-1} \log_k^{\circ j}(n)}{\prod_{j=2}^{i-1} \log_k^{\circ j}(n^\gamma)} = 1.$$

□

7.3 Closed words

7.3.1 Lower bound

We first state a useful lemma from a paper of Nicholson and Rampersad [79].

Lemma 69 (Nicholson and Rampersad [79]). *Let $k \geq 2$ be an integer. Let t be the unique integer such that*

$$\frac{\ln k}{k-1}k^t \leq n - t < \frac{\ln k}{k-1}k^{t+1}.$$

Let u be a length-t word. There exist constants N_0 and d such that for $n - t > N_0$ we have

$$B_k(n, u) \geq d\frac{k^n}{n^2}.$$

We now use the previous lemma to prove Theorem 65 (a).

Proof of Theorem 65 (a). The number $C_k(n, t)$ of length-n words closed by a length-t word is clearly equal to the sum, over all length-t words u, of the number $B_k(n, u)$ of length-n words closed by u. Thus we have that

$$C_k(n, t) = \sum_{|u|=t} B_k(n, u).$$

Let $t = \lfloor \log_k(n - t) + \log_k(k - 1) - \log_k(\ln k) \rfloor$. By Lemma 69 there exist constants N_0 and d such that for $n - t > N_0$ we have $B_k(n, u) \geq dk^n/n^2$. Clearly $t \leq \log_k(n) + 1$ for all $n \geq 1$. Since t is asymptotically much smaller than n, there exists a constant $N > N_0$ such that $n - t > N_0$ for all $n > N$. Thus for $n > N$ we have

$$\begin{aligned} C_k(n) \geq C_k(n, t) &= \sum_{|u|=t} B_k(n, u) \geq \sum_{|u|=t} d\frac{k^n}{n^2} = k^t\left(d\frac{k^n}{n^2}\right) \\ &= dk^{\lfloor \log_k(n-t)+\log_k(k-1)-\log_k(\ln k)\rfloor}\frac{k^n}{n^2} \geq d_0 k^{\log_k(n-t)+\log_k(k-1)-\log_k(\ln k)}\frac{k^n}{n^2} \\ &\geq d_1(n-t)\frac{k^n}{n^2} \geq d_1(n - \log_k(n) - 1)\frac{k^n}{n^2} \geq c\frac{k^n}{n} \end{aligned}$$

for some constant $c > 0$. □

7.3.2 Upper bound

Before we proceed with upper bounding $C_k(n)$, we briefly outline the direction of the proof. First, we begin by bounding $C_k(n,t)$ for $t < n/2$ and $t \geq n/2$. We show that for $t < n/2$, the number of length-n words closed by a particular length-t word u is bounded by the number of words of length $n-2t$ that do not have 0^t as a factor. For $t \geq n/2$ we prove that $C_k(n,t)$ is negligibly small. Next, we prove upper bounds on the number of words that do not have 0^t as a factor, allowing us to finally bound $C_k(n)$.

Lemma 70. *Let n, t, and k be integers such that $n \geq 2t \geq 2$ and $k \geq 2$. Let u be a length-t word. Then*

$$B_k(n,u) \leq A_k(n-2t, 0^t).$$

Proof. Recall that $B_k(n,u)$ is the number of length-n words that are closed by the word u.

Let w be a length-n word closed by u where $|w| = n \geq 2t = 2|u|$. Then we can write $w = uvu$ where v does not contain u as a factor. This immediately implies that $B_k(n,u) \leq A_k(n-2t,u)$. But from a result of Guibas and Odlyzko [52, Section 7], we have that if $f_{a(u)}(2) > f_{a(v)}(2)$ for words u, v, then $A_k(m,u) \geq A_k(m,v)$ for all $m \geq 1$. The auto-correlation polynomial only has 0 or 1 as coefficients, depending on the 1's and 0's in the auto-correlation. Thus the auto-correlation p that maximizes $f_p(2)$ is clearly $p = 1^t$. The words that achieve this auto-correlation are words of the form a^t where $a \in \Sigma_k$. Therefore we have

$$B_k(n,u) \leq A_k(n-2t,u) \leq A_k(n-2t,0^t).$$

□

Lemma 71. *Let n, t, and k be integers such that $n \geq 2t \geq 2$ and $k \geq 2$. Then*

$$C_k(n,t) \leq k^t A_k(n-2t,0^t).$$

Proof. The number $C_k(n,t)$ of length-n words closed by a length-t word is equal to the sum, over all length-t words u, of the number $B_k(n,u)$ of length-n words closed by u. Thus we have that

$$C_k(n,t) = \sum_{|u|=t} B_k(n,u).$$

By Lemma 70 we have that $B_k(n,v) \leq A_k(n-2t,0^t)$ for all length-t words v. Therefore

$$C_k(n,t) = \sum_{|u|=t} B_k(n,u) \leq \sum_{|u|=t} A_k(n-2t,0^t) \leq k^t A_k(n-2t,0^t).$$

□

Corollary 72. *Let $n \geq 1$ and $k \geq 2$ integers. Then*

$$C_k(n) \leq \sum_{t=1}^{\lfloor n/2 \rfloor} k^t A_k(n-2t, 0^t) + nk^{\lceil n/2 \rceil}.$$

Proof. It follows from Lemma 71 that

$$C_k(n) = \sum_{t=1}^{n-1} C_k(n,t) \leq \sum_{t=1}^{\lfloor n/2 \rfloor} k^t A_k(n-2t, 0^t) + \sum_{t=\lfloor n/2 \rfloor + 1}^{n-1} C_k(n,t).$$

Now we show that

$$\sum_{t=\lfloor n/2 \rfloor + 1}^{n-1} C_k(n,t) \leq nk^{\lceil n/2 \rceil}.$$

Let $w = w_0 w_1 \cdots w_{n-1}$ be a word of length n that is closed by a word u of length $t > \lfloor n/2 \rfloor$. Then $w = ux = yu$ for some words x, y. So $w_i = w_{i+(n-t)}$ for all i, $0 \leq i < t$. This implies that $w = v^i v'$ where v is the length-$(n-t)$ prefix of w, $i = \lfloor n/|v| \rfloor$, and v' is the length-$(n - i|v|)$ prefix of v. Since $t > \lfloor n/2 \rfloor$, we have that $n - t < \lceil n/2 \rceil$. We see that w is fully determined by the word v. So since $|v| < \lceil n/2 \rceil$, we have $C_k(n,t) \leq k^{\lceil n/2 \rceil}$. Thus

$$\sum_{t=\lfloor n/2 \rfloor + 1}^{n-1} C_k(n,t) \leq \sum_{t=\lfloor n/2 \rfloor + 1}^{n-1} k^{\lceil n/2 \rceil} \leq nk^{\lceil n/2 \rceil}.$$

□

Lemma 73. *Let $n \geq 0$, $t \geq 1$, and $k \geq 2$ be integers. Then*

$$A_k(n, 0^t) = \begin{cases} k^n, & \text{if } n < t; \\ (k-1) \sum_{i=1}^{t} A_k(n-i, 0^t), & \text{if } n \geq t. \end{cases}$$

Proof. If $n < t$, then any length-n word is shorter than 0^t, and thus cannot contain 0^t as a factor. So $A_k(n, 0^t) = k^n$.

Suppose $n \geq t$. Let w be a length-n word that does not contain 0^t as a factor. Let us look at the symbols that w ends in. Since w does not contain 0^t, we have that w ends in anywhere from 0 to $t-1$ zeroes. So w is of the form $w = w'b0^i$ where i is an integer with $0 \leq i \leq t-1$, $b \in \Sigma_k - \{0\}$, and w' is a length-$(n-i-1)$ word that does not contain 0^t

as a factor. There are $k-1$ choices for b, and $A_k(n-i-1,0^t)$ choices for w'. So there are $(k-1)A_k(n-i-1,0^t)$ words of the form $w'b0^i$. Summing over all possible i gives

$$A_k(n,0^t) = (k-1)\sum_{i=1}^{t} A_k(n-i,0^t).$$

□

Corollary 74. *Let $n \geq 0$, $t \geq 1$, and $k \geq 2$ be integers. Then*

$$A_k(n,0^t) = \begin{cases} k^n, & \text{if } n < t; \\ k^n - 1, & \text{if } n = t; \\ kA_k(n-1,0^t) - (k-1)A_k(n-t-1,0^t), & \text{if } n > t. \end{cases}$$

Proof. Compute $A_k(n,0^t) - A_k(n-1,0^t)$ with the recurrence from Lemma 73 and the result follows. □

Corollary 75. *Let $n \geq 0$, $t \geq 1$, and $k \geq 2$ be integers. Then*

$$A_k(n,0^t) = \begin{cases} k^n, & \text{if } n < t; \\ k^{n-t}(k^t-1) - (n-t)k^{n-t-1}(k-1), & \text{if } t \leq n \leq 2t; \\ kA_k(n-1,0^t) - (k-1)A_k(n-t-1,0^t), & \text{if } n > 2t. \end{cases}$$

Since $(A_k(n,0^t))_n$ satisfies a linear recurrence, we know that the asymptotic behaviour of $A_k(n,0^t)$ is determined by the real roots of the polynomial $x^{t+1} - kx^t + k - 1 = 0$. We use this fact to find an upper bound for $A_k(n,0^t)$.

Theorem 76. *Let $t \geq 1$ and $k \geq 2$ be integers. Let*

$$\beta_k(t) = k - (k-1)k^{-t-1}.$$

Then $\beta_k(t) \geq k - (k-1)\beta_k(t)^{-t}$.

Proof. Since $\beta_k(t) \leq k$, we have that $\beta_k(t)^{-t} \geq k^{-t} \geq k^{-t-1}$. This implies that

$$\beta_k(t) = k - (k-1)k^{-t-1} \geq k - (k-1)\beta_k(t)^{-t}.$$

□

Lemma 77. *Let $k, t \geq 2$ be integers. Let n be an integer such that $2t \leq n \leq 3t$. Then $A_k(n, 0^t) \leq \beta_k(t)^n$.*

Proof. The proof is by induction on n. By Corollary 75 we have that

$$A_k(n, 0^t) = k^{n-t}(k^t - 1) - (n-t)k^{n-t-1}(k-1)$$

for $t \leq n \leq 2t$. Let $\gamma(t) = (k-1)k^{-t-1}$.

Suppose, for the base case, that $n = 2t$. Then

$$\begin{aligned} A_k(2t, 0^t) &= k^t(k^t - 1) - tk^{t-1}(k-1) = k^{2t} - k^{t-2}(k^2 + tk(k-1)) \\ &= k^{2t} - \gamma(t)k^{2t-1}\frac{(k^2 + tk(k-1))}{k-1} \\ &\leq k^{2t} - \gamma(t)tk^{2t-1}. \end{aligned}$$

Clearly $\gamma(t) \leq 6/t$ for all $t \geq 2$, so $A_k(2t) \leq k^{2t} - \gamma(t)tk^{2t-1} \leq (k - \gamma(t))^{2t} = \beta_k(t)^{2t}$.

Suppose that $2t < n \leq 3t$. Furthermore let $n = 2t + i + 1$ where i is an integer such that $0 \leq i < t$. Notice that $A_k(n - t - 1, 0^t) = A_k(t + i, 0^t) = k^i(k^t - 1) - ik^{i-1}(k-1)$. Then

$$\begin{aligned} A_k(2t + i + 1, 0^t) &= kA_k(2t + i, 0^t) - (k-1)A_k(t + i, 0^t) \\ &\leq k(k - \gamma(t))^{2t+i} - (k-1)(k^i(k^t - 1) - ik^{i-1}(k-1)) \\ &= (k - \gamma(t))^{2t+i+1} + \gamma(t)(k - \gamma(t))^{2t+i} - (k-1)(k^i(k^t - 1) - ik^{i-1}(k-1)) \\ &= \beta_k(t)^{2t+i+1} + \gamma(t)\beta_k(t)^{2t+i} - (k-1)(k^i(k^t - 1) - ik^{i-1}(k-1)). \end{aligned}$$

To prove the desired bound, namely that $A_k(2t + i + 1, 0^t) \leq \beta_k(t)^{2t+i+1}$, it is sufficient to show that $\beta_k(t)^{2t+i} \leq \gamma(t)^{-1}(k-1)(k^i(k^t - 1) - ik^{i-1}(k-1))$. We begin by upper bounding $\beta_k(t)^{2t+i}$ with Lemma 67. We have

$$\begin{aligned} \beta_k(t)^{2t+i} &\leq k^{2t+i} - \gamma(t)(2t+i)k^{2t+i-1} + \frac{1}{2}\gamma(t)^2(2t+i)(2t+i-1)k^{2t+i-2} \\ &\leq k^{2t+i} - 2(k-1)tk^{t+i-2} + \frac{9}{2}(k-1)^2t^2k^{i-4} \\ &\leq k^{2t+i+1} - (k-1)k^{2t+i} - 2(k-1)tk^{t+i-2} + \frac{9}{2}(k-1)^2t^2k^{i-4} \\ &= k^{2t+i+1} - k^{t+i}\Big((k-1)k^t + 2(k-1)tk^{-2} - \frac{9}{2}(k-1)^2t^2k^{-t-4}\Big). \end{aligned} \tag{7.2}$$

It is easy to verify that $(k-1)k^t \geq k + t(k-1)$ and $2(k-1)tk^{-2} - \frac{9}{2}(k-1)^2t^2k^{-t-4} \geq 0$ for all $t \geq 2$. Thus, continuing from (7.2), we have

$$\begin{aligned}\beta_k(t)^{2t+i} &\leq k^{2t+i+1} - k^{t+i}(k + t(k-1)) \leq k^{2t+i+1} - k^{t+i}(k + i(k-1)) \\ &= \frac{k^{t+1}}{k-1}(k-1)(k^{t+i} - k^i - ik^{i-1}(k-1)) \\ &= \gamma(t)^{-1}(k-1)(k^i(k^t-1) - ik^{i-1}(k-1)).\end{aligned}$$

□

Lemma 78. *Let n, t, and k be integers such that $n \geq 2t \geq 4$ and $k \geq 2$. Then $A_k(n, 0^t) \leq \beta_k(t)^n$.*

Proof. The proof is by induction on n. The base case, when $2t \leq n \leq 3t$, is taken care of in Lemma 77.

Suppose $n > 3t$. Then

$$A_k(n, 0^t) = (k-1)\sum_{i=1}^{t} A_k(n-i, 0^t) \leq (k-1)\sum_{i=1}^{t} \beta_k(t)^{n-i} = (k-1)\frac{\beta_k(t)^n - \beta_k(t)^{n-t}}{\beta_k(t) - 1}.$$

By Theorem 76, we have that $\beta_k(t) - 1 \geq (k-1) - (k-1)\beta_k(t)^{-t}$. Therefore

$$A_k(n, 0^t) \leq (k-1)\frac{\beta_k(t)^n - \beta_k(t)^{n-t}}{\beta_k(t) - 1} = \beta_k(t)^n \frac{(k-1) - (k-1)\beta_k(t)^{-t}}{\beta_k(t) - 1} \leq \beta_k(t)^n.$$

□

Proof of Theorem 65 (b). First notice that $A_k(n, 0) = (k-1)^n$, since $A_k(n, 0)$ is just the number of length-n words that do not contain 0.

Let N' be a positive integer such that the following inequalities hold for all $n > N'$.

$$
\begin{aligned}
C_k(n) &\le \sum_{t=2}^{\lfloor n/2 \rfloor} k^t A_k(n-2t, 0^t) + kA_k(n-2,0) + nk^{\lceil n/2 \rceil} \\
&\le \sum_{t=2}^{\lfloor n/2 \rfloor} k^t \beta_k(t)^{n-2t} + k(k-1)^{n-2} + nk^{\lceil n/2 \rceil} \\
&\le \sum_{t=2}^{\lfloor n/2 \rfloor} k^t \left(k - \frac{k-1}{k^{t+1}}\right)^{n-2t} + d_2 \frac{k^n}{n} = k^n \sum_{t=2}^{\lfloor n/2 \rfloor} \frac{1}{k^t}\left(1 - \frac{k-1}{k^{t+2}}\right)^{n-2t} + d_2 \frac{k^n}{n} \\
&\le k^n \left(\sum_{t=2}^{\lfloor \log_k n \rfloor} \frac{1}{k^t}\left(1 - \frac{k-1}{k^{t+2}}\right)^{n-2t} + \sum_{t=\lfloor \log_k n \rfloor + 1}^{\lfloor n/2 \rfloor} \frac{1}{k^t} \right) + d_2 \frac{k^n}{n} \\
&\le k^n \left(\sum_{t=2}^{\lfloor \log_k n \rfloor} \frac{1}{k^t}\left(1 - \frac{k-1}{k^{t+2}}\right)^{n-2\lfloor \log_k n \rfloor} + \frac{d_3}{n} \right) + d_2 \frac{k^n}{n} \\
&\le k^n \sum_{t=2}^{\lfloor \log_k n \rfloor} \frac{1}{k^t}\left(1 - \frac{k-1}{k^{t+2}}\right)^{n/2} + d_4 \frac{k^n}{n}. \qquad (7.3)
\end{aligned}
$$

Now we bound the sum in (7.3). Let $h(x) = (1-(k-1)k^{-2}x)^{n/2}$. Notice that $h(x)$ is monotonically decreasing on the interval $x \in (0,1)$. So for $k^{-t-1} \le x \le k^{-t}$ we have that $h(x) \ge h(k^{-t})$. Thus

$$
\frac{1}{k^t}\left(1 - \frac{k-1}{k^{t+2}}\right)^{n/2} \le \frac{k-1}{k^t}\left(1 - \frac{k-1}{k^{t+2}}\right)^{n/2} \le k\left(\left(\frac{1}{k^t} - \frac{1}{k^{t+1}}\right)h(k^{-t})\right) \le k \int_{k^{-t-1}}^{k^{-t}} h(x)\,dx.
$$

Going back to (7.3) we have

$$
C_k(n) \le k^n \sum_{t=2}^{\lfloor \log_k n \rfloor} k \int_{k^{-t-1}}^{k^{-t}} h(x)\,dx + d_4 \frac{k^n}{n} \le k^{n+1} \int_0^1 h(x)\,dx + d_4 \frac{k^n}{n}.
$$

Evaluating and bounding the definite integral, we have

$$
\begin{aligned}
\int_0^1 h(x)\,dx &= -\frac{k^2}{k-1}\left[\frac{(1-(k-1)k^{-2}x)^{n/2+1}}{n/2+1}\right]_{x=0}^{x=1} \\
&= -\frac{k^2}{k-1}\left(\frac{(1-(k-1)k^{-2})^{n/2+1} - 1}{n/2+1}\right)
\end{aligned}
$$

$$\leq d_5\left(\frac{1-(1-(k-1)k^{-2})^{n/2+1}}{n/2+1}\right) \leq d_5\frac{1}{n/2+1} \leq \frac{d_6}{n}.$$

Putting everything together, we have that

$$C_k(n) \leq k^{n+1}\int_0^1 h(x)\,dx + d_4\frac{k^n}{n} \leq d_6\frac{k^{n+1}}{n} + d_4\frac{k^n}{n} \leq c'\frac{k^n}{n}$$

for some constant $c' > 0$. □

7.4 Privileged words

7.4.1 Lower bound

In this section we provide a family of lower bounds for the number of length-n privileged words. We use induction to prove these bounds. The basic idea is that we start with the lower bound by Nicholson and Rampersad, and then use it to bootstrap ourselves to better and better lower bounds.

Proof of Theorem 66 (a). The proof is by induction on j. Let

$$t = \lfloor \log_k(n-t) + \log_k(k-1) - \log_k(\ln k)\rfloor.$$

We clearly have $0 \leq t \leq \log_k(n)+1$ for all $n \geq 1$. Let u be a length-t privileged word. By Lemma 69 we have that there exist constants N_0 and c_0 such that $P_k(n) \geq B_k(n,u) \geq c_0\frac{k^n}{n^2}$ for all $n > N_0$. So the base case, when $j = 0$, is taken care of.

Suppose $j > 0$. By induction we have that there exist constants N_{j-1} and c_{j-1} such that

$$P_k(n) \geq c_{j-1}\frac{k^n}{n\log_k^{\circ j-1}(n)\prod_{i=1}^{j-1}\log_k^{\circ i}(n)}$$

for all $n > N_{j-1}$. By Lemma 69 we have

$$P_k(n) \geq P_k(n,t) \geq \sum_{\substack{|u|=t \\ u \text{ privileged}}} B_k(n,u) \geq \sum_{\substack{|u|=t \\ u \text{ privileged}}} d\frac{k^n}{n^2} = dP_k(t)\frac{k^n}{n^2}.$$

for $n > N_0$. Since $t \leq \log_k(n) + 1$, we have that $\frac{1}{\log_k^{\circ i}(t)} \geq \frac{1}{\log_k^{\circ i}(\log_k(n)+1)}$ for all $i \geq 0$. Thus continuing from above we have

$$\begin{aligned}
P_k(n) &\geq dc_{j-1} \frac{k^t}{t \log_k^{\circ j-1}(t) \prod_{i=1}^{j-1} \log_k^{\circ i}(t)} \frac{k^n}{n^2} \geq d_7 \frac{k^{\log_k(n-t)+\log_k(k-1)-\log_k(\ln k)}}{t \log_k^{\circ j-1}(t) \prod_{i=1}^{j-1} \log_k^{\circ i}(t)} \frac{k^n}{n^2} \\
&\geq d_8 \frac{1}{t \log_k^{\circ j-1}(t) \prod_{i=1}^{j-1} \log_k^{\circ i}(t)} \frac{k^n}{n} \\
&\geq d_9 \frac{1}{(\log_k(n)+1) \log_k^{\circ j-1}(\log_k(n)+1) \prod_{i=1}^{j-1} \log_k^{\circ i}(\log_k(n)+1)} \frac{k^n}{n} \\
&\geq c_j \frac{k^n}{n \log_k^{\circ j}(n) \prod_{i=1}^{j} \log_k^{\circ i}(n)}
\end{aligned}$$

for all $n > N_j$ where $N_j > \max(N_0, N_{j-1})$. □

7.4.2 Upper bound

Theorem 65 (b) immediately implies that $P_k(n) \in O(\frac{k^n}{n})$. This bound improves on the existing bound on privileged words, but it does not show that $P_k(n)$ and $C_k(n)$ behave differently asymptotically. We show that $P_k(n)$ is much smaller than $C_k(n)$ asymptotically by proving upper bounds on $P_k(n)$ that show $P_k(n) \in o(\frac{k^n}{n})$.

Lemma 79. *Let n, t, and k be integers such that $n \geq 2t \geq 2$ and $k \geq 2$. Then*

$$P_k(n,t) \leq P_k(t) A_k(n-2t, 0^t).$$

Proof. The number of length-n privileged words closed by a length-t privileged word is equal to the sum, over all length-t privileged words u, of the number $B_k(n,u)$ of length-n words closed by u. Thus we have that

$$P_k(n,t) = \sum_{\substack{|u|=t \\ u \text{ privileged}}} B_k(n,u).$$

By Lemma 70 we have that $B_k(n,v) \leq A_k(n-2t, 0^t)$ for all length-t words v. Therefore

$$P_k(n,t) = \sum_{\substack{|u|=t \\ u \text{ privileged}}} B_k(n,u) \leq \sum_{\substack{|u|=t \\ u \text{ privileged}}} A_k(n-2t, 0^t) \leq P_k(t) A_k(n-2t, 0^t).$$

□

Proof of Theorem 66 (b). For $n \geq 2t$ we can use Lemma 79 to bound $P_k(n,t)$. But for $n < 2t$, we can use Corollary 72 and the fact that $P_k(n,t) \leq C_k(n,t)$. We get

$$P_k(n) = \sum_{t=1}^{n-1} P_k(n,t) \leq \sum_{t=1}^{\lfloor n/2 \rfloor} P_k(t) A_k(n-2t, 0^t) + nk^{\lceil n/2 \rceil}.$$

The proof is by induction on j. The base case, when $j = 0$, is taken care of by Theorem 65 (b).

Suppose $j > 0$. Then there exist constants N'_{j-1} and c'_{j-1} such that

$$P_k(n) \leq c'_{j-1} \frac{k^n}{n \prod_{i=1}^{j-1} \log_k^{\circ i}(n)}$$

for all $n > N'_{j-1}$. We now bound $P_k(n)$. First we let $N'_j > N'_{j-1}$ be a constant such that the following inequalities hold for all $n > N'_j$. We have

$$\begin{aligned}
P_k(n) &\leq \sum_{t=1}^{\lfloor n/2 \rfloor} P_k(t) A_k(n-2t, 0^t) + nk^{\lceil n/2 \rceil} \\
&\leq \sum_{t=N'_j+1}^{\lfloor n/2 \rfloor} c'_{j-1} \frac{k^t}{t \prod_{i=1}^{j-1} \log_k^{\circ i}(t)} \beta_k(t)^{n-2t} + \sum_{t=1}^{N'_j} P_k(t) A_k(n-2t, 0^t) + nk^{\lceil n/2 \rceil} \\
&\leq \sum_{t=N'_j+1}^{\lfloor n/2 \rfloor} c'_{j-1} \frac{k^t}{t \prod_{i=1}^{j-1} \log_k^{\circ i}(t)} \left(k - \frac{k-1}{k^{t+1}}\right)^{n-2t} + d_{10} \sum_{t=2}^{N'_j} \left(k - \frac{k-1}{k^{t+1}}\right)^{n-2t} + d_{11} \frac{k^n}{n^2} \\
&\leq c'_{j-1} k^n \sum_{t=N'_j+1}^{\lfloor n/2 \rfloor} \frac{1}{k^t t \prod_{i=1}^{j-1} \log_k^{\circ i}(t)} \left(1 - \frac{k-1}{k^{t+2}}\right)^{n-2t} + d_{12} \frac{k^n}{n^2} \\
&\leq c'_{j-1} k^n \left(d_{13} \sum_{t=N'_j+1}^{\lfloor \log_k(n) \rfloor} \frac{1}{k^t t \prod_{i=1}^{j-1} \log_k^{\circ i}(t)} \left(1 - \frac{k-1}{k^{t+2}}\right)^{n/2} + \sum_{t=\lfloor \log_k(n) \rfloor + 1}^{\lfloor n/2 \rfloor} \frac{1}{k^t t \prod_{i=1}^{j-1} \log_k^{\circ i}(t)} \right) + d_{12} \frac{k^n}{n^2} \\
&\leq c'_{j-1} k^n \left(d_{13} \sum_{t=N'_j+1}^{\lfloor \log_k(n) \rfloor} \frac{1}{k^t t \prod_{i=1}^{j-1} \log_k^{\circ i}(t)} \exp\left(\frac{n}{2} \ln\left(1 - \frac{k-1}{k^{t+2}}\right)\right) \right. \\
&\qquad \left. + \sum_{t=\lfloor \log_k(n) \rfloor + 1}^{\infty} \frac{1}{k^t t \prod_{i=1}^{j-1} \log_k^{\circ i}(t)} \right) + d_{12} \frac{k^n}{n^2} \qquad (7.4)
\end{aligned}$$

The sum on line (7.4) is clearly convergent. We have

$$\sum_{t=\lfloor \log_k(n)\rfloor+1}^{\infty} \frac{1}{k^t t \prod_{i=1}^{j-1} \log_k^{\circ i}(t)} \leq \frac{1}{(\lfloor \log_k(n)\rfloor+1)\prod_{i=1}^{j-1} \log_k^{\circ i}(\lfloor \log_k(n)\rfloor+1)} \sum_{t=\lfloor \log_k(n)\rfloor+1}^{\infty} \frac{1}{k^t}$$
$$\leq d_{14} \frac{1}{\log_k(n)\prod_{i=1}^{j-1} \log_k^{\circ i}(\log_k(n))} \frac{1}{n} \leq d_{14} \frac{1}{n\prod_{i=1}^{j} \log_k^{\circ i}(n)}.$$

Now we upper bound the sum

$$D_n = \sum_{t=N'_j+1}^{\lfloor \log_k(n)\rfloor} \frac{1}{k^t t \prod_{i=1}^{j-1} \log_k^{\circ i}(t)} \exp\left(\frac{n}{2} \ln\left(1 - \frac{k-1}{k^{t+2}}\right)\right).$$

It is well known that $\ln(1-x) \leq -x$ for $|x| < 1$. Thus, letting $\alpha = \frac{k-1}{2k^2}$, we have

$$\exp\left(\frac{n}{2} \ln\left(1 - \frac{k-1}{k^{t+2}}\right)\right) \leq \exp\Big(-\alpha \frac{n}{k^t}\Big).$$

We reverse the order of the series, by letting $t = \lfloor \log_k(n)\rfloor - t + N'_j + 1$. We also shift the index of the series down by $N'_j + 1$. We have

$$\begin{aligned} D_n &= \sum_{t=0}^{\lfloor \log_k(n)\rfloor - N'_j - 1} \frac{1}{k^{\lfloor \log_k(n)\rfloor - t}(\lfloor \log_k(n)\rfloor - t)\prod_{i=1}^{j-1} \log_k^{\circ i}(\lfloor \log_k(n)\rfloor - t)} \exp\Big(-\alpha \frac{n}{k^{\lfloor \log_k(n)\rfloor - t}}\Big) \\ &\leq d_{15} \sum_{t=0}^{\lfloor \log_k(n)\rfloor - N'_j - 1} \frac{k^t}{n(\log_k(n) - t)\prod_{i=1}^{j-1} \log_k^{\circ i}(\log_k(n) - t)} \exp\left(-\alpha k^t\right) \\ &\leq d_{15} \frac{1}{n\prod_{i=1}^{j} \log_k^{\circ i}(n)} \sum_{t=0}^{\lfloor \log_k(n)\rfloor - N'_j - 1} \frac{k^t}{\prod_{i=0}^{j-1} \frac{\log_k^{\circ i}(\log_k(n)-t)}{\log_k^{\circ i+1}(n)}} \exp\left(-\alpha k^t\right). \end{aligned} \tag{7.5}$$

Suppose β is a positive constant strictly between 0 and 1 such that $\beta \log_k(n)$ is an integer and $\beta \log_k(n) < \lfloor \log_k(n) \rfloor - N'_j - 1$. If $t \leq \beta \log_k(n)$, then

$$\frac{\log_k^{\circ i}(\log_k(n) - t)}{\log_k^{\circ i+1}(n)} \geq \frac{\log_k^{\circ i+1}(n^{1-\beta})}{\log_k^{\circ i+1}(n)} \geq d'_i$$

for some $d'_i > 0$ by Lemma 68. If $t > \beta \log_k(n)$, then

$$\frac{\log_k^{\circ i}(\log_k(n) - t)}{\log_k^{\circ i+1}(n)} \geq \frac{\log_k^{\circ i}(N'_j + 1)}{\log_k^{\circ i+1}(n)}.$$

We split up the sum in D_n in two parts. One sum with $t \leq \beta \log_k(n)$ and one with $t > \beta \log_k(n)$. Continuing from (7.5) we get

$$\begin{aligned}
&\leq d_{15} \frac{1}{n \prod_{i=1}^{j} \log_k^{\circ i}(n)} \left(\sum_{t=1}^{\beta \log_k(n)} \frac{k^t}{\prod_{i=0}^{j-1} d'_i} \exp\left(-\alpha k^t\right) + \prod_{i=0}^{j-1} \left(\frac{\log_k^{\circ i+1}(n)}{\log_k^{\circ i}(N'_j + 1)} \right) \sum_{t=\beta \log_k(n)+1}^{\lfloor \log_k(n) \rfloor - N'_j - 1} k^t \exp\left(-\alpha k^t\right) \right) \\
&\leq d_{15} \frac{1}{n \prod_{i=1}^{j} \log_k^{\circ i}(n)} \left(d_{16} \sum_{t=1}^{\infty} t \exp\left(-\alpha t\right) + d_{17} \prod_{i=1}^{j} \log_k^{\circ i}(n) \sum_{t=kn^\beta}^{\infty} t \exp\left(-\alpha t\right) \right).
\end{aligned}$$

The first and second sum are both clearly convergent. It is also easy to show that both of them can be bounded by a constant multiplied by the first term. Thus we have that

$$D_n \leq d_{15} \frac{1}{n \prod_{i=1}^{j} \log_k^{\circ i}(n)} \left(d_{18} + d_{19} \prod_{i=1}^{j} \log_k^{\circ i}(n) \frac{kn^\beta}{\exp\left(\alpha k n^\beta\right)} \right) \leq d_{20} \frac{1}{n \prod_{i=1}^{j} \log_k^{\circ i}(n)}.$$

Putting everything together and continuing from line (7.4), we get

$$P_k(n) \leq c' k^n \left(d_{13} D_n + d_{14} \frac{1}{n \prod_{i=1}^{j} \log_k^{\circ i}(n)} \right) + d_{12} \frac{k^n}{n^2} \leq c'_j \frac{k^n}{n \prod_{i=1}^{j} \log_k^{\circ i}(n)}$$

for some constant $c'_j > 0$. □

Chapter 8

Mutual borders and overlaps

8.1 Introduction

Let $\mathcal{U}_n^k$ denote the set of length-n unbordered words over a k-letter alphabet. It is well known [80] that the sequence $u_n = |\mathcal{U}_n^k|$ is defined by the recurrence

$$u_n = \begin{cases} 1, & \text{if } n = 0; \\ ku_{n-1} - u_{n/2}, & \text{if } n > 0 \text{ is even}; \\ ku_{n-1}, & \text{if } n \text{ is odd.} \end{cases}$$

In the same paper by Nielsen, he showed that the limit $\lim_{n\to\infty} u_n/k^n$ exists. In particular, he showed that for $k = 2$ there are $(c + o(1)) \cdot 2^n$ unbordered binary words, where $c \approx 0.267786$. As we saw in Section 2.4, the notion of a word being unbordered or bordered can naturally be generalized to pairs of words. In this chapter we prove results similar to Nielsen's for these kinds of pairs of words.[1]

Let u and v be words of length m and n, respectively. Let w be a non-empty word. In this chapter we write (u, v) to refer to an ordered pair of words. Recall some definitions from Section 2.4.

- We say that (u, v) has a *right-border* w (resp., *left-border*) if w is a non-empty proper suffix (resp., prefix) of u that is a proper prefix (resp., suffix) of v.

[1] These results appear in [43].

- The pair (u, v) is said to be *mutually bordered* if (u, v) has both a right-border and a left-border.
- If (u, v) has neither a right-border nor a left-border, then (u, v) is said to be *mutually unbordered*.
- The pair (u, v) is said to be *right-bordered* (resp., *left-bordered*) if (u, v) has a right-border (resp., *left-border*) but not a left-border (resp., *right-border*).[2]

See Example 16 for examples illustrating these different definitions.

- Let $M_k(m, n)$ denote the number of mutually bordered pairs of words (u, v).
- Let $R_k(m, n)$ denote the number of right-bordered pairs of words (u, v).
- Let $U_k(m, n)$ denote the number of mutually unbordered pairs of words (u, v).

See Tables 8.1, 8.2, and 8.3 for sample values of $M_k(m, n)$, $R_k(m, n)$, and $U_k(m, n)$ for m, n where $1 \leq m, n \leq 8$.

In this chapter we are primarily concerned with pairs of equal length words (i.e., the case where $m = n$). So we define $M_k(n) = M_k(n, n)$, and we define $R_k(n)$ and $U_k(n)$ similarly (see Table 8.4 for some sample values). The main results of this chapter are Corollary 84, Theorem 87, Theorem 88, and Theorem 92. In Corollary 84 we bound the sum of the length of the shortest left-border and right-border of a pair of words. In Theorem 87 we give recurrences for $M_k(n), R_k(n)$, and $U_k(n)$. Then in Theorem 88 we prove that the limit $\lim_{n\to\infty} M_k(n)/k^{2n}$ exists. Finally, in Theorem 92 we show that the expected shortest right-border and left-border of a pair of equal-length words is $O(1)$.

8.2 Number of mutually bordered pairs

In this section we enumerate the number of mutually bordered pairs of words $M_k(n)$.

- Let $\text{so}(u, v)$ denote the shortest right-border of (u, v), and let $\text{so}(u, v) = \epsilon$ if (u, v) does not have a right-border. By definition we have that $\text{so}(v, u)$ is the shortest left-border of (v, u), and $\text{so}(v, u) = \epsilon$ if (u, v) does not have a left-border.

[2]We could have defined left-borders and right-borders to refer to ordinary non-empty prefixes and suffixes without specifying they be proper. But since a border is defined as a non-empty proper prefix and suffix of a word, we decided to keep the definition analogous.

Table 8.1: Some values of $M_2(m, n)$ for m, n where $1 \leq m, n \leq 8$.

m \ n	1	2	3	4	5	6	7	8
1	0	0	0	0	0	0	0	0
2	0	4	8	16	32	64	128	256
3	0	8	26	50	100	200	400	800
4	0	16	50	124	242	484	968	1936
5	0	32	100	242	524	1036	2070	4142
6	0	64	200	484	1036	2154	4280	8554
7	0	128	400	968	2070	4280	8706	17354
8	0	256	800	1936	4142	8554	17354	34996

Table 8.2: Some values of $R_2(m, n)$ for m, n where $1 \leq m, n \leq 8$.

m \ n	1	2	3	4	5	6	7	8
1	0	0	0	0	0	0	0	0
2	0	4	8	16	32	64	128	256
3	0	8	14	30	60	120	240	480
4	0	16	30	52	110	220	440	880
5	0	32	60	110	204	420	842	1682
6	0	64	120	220	420	806	1640	3286
7	0	128	240	440	842	1640	3214	6486
8	0	256	480	880	1682	3286	6486	12844

- Let $\mathrm{lso}(v, u)$ be the length of the shortest right-border of (u, v), and let $\mathrm{lso}(u, v) = 0$ if (u, v) does not have a right-border. Again we have that by definition $\mathrm{lso}(v, u)$ is the length of the shortest left-border of (u, v), and $\mathrm{lso}(v, u) = 0$ if (u, v) does not have a left-border.

Example 80. The pair of binary words $(u, v) = (1000101, 0110001)$ has one right-border and two left-borders. The word 01 is a right-border of the pair. The words 1 and 10001 are left-borders of the pair. The shortest right-border is 01 and is of length 2. The shortest left-border is 1 and is of length 1. So $\mathrm{so}(u, v) = 01$, $\mathrm{lso}(u, v) = 2$, $\mathrm{so}(v, u) = 1$, and $\mathrm{lso}(v, u) = 1$.

Since the concept of a pair of words being mutually bordered is similar to the concept of a single word being bordered, it is natural to assume that the insights used to count

Table 8.3: Some values of $U_2(m, n)$ for m, n where $1 \leq m, n \leq 8$.

m \ n	1	2	3	4	5	6	7	8
1	4	8	16	32	64	128	256	512
2	8	4	8	16	32	64	128	256
3	16	8	10	18	36	72	144	288
4	32	16	18	28	50	100	200	400
5	64	32	36	50	92	172	342	686
6	128	64	72	100	172	330	632	1258
7	256	128	144	200	342	632	1250	2442
8	512	256	288	400	686	1258	2442	4852

bordered words might be useful to also count pairs of mutually bordered words. Let u be a bordered word. Let v be the shortest border of u. The key ideas used to count length-n bordered words are that for bordered words u, the shortest border v is unbordered, and $|v| \leq n/2$. Combining these facts we get the following formula for the number of length-n bordered words over a k-letter alphabet:

$$\sum_{i=1}^{\lfloor n/2 \rfloor} u_i \cdot k^{n-2i}.$$

The basic idea we will use to count mutually bordered pairs of words (u, v) is to, analogously to bounding the length of the shortest border, put a bound on $\mathrm{lso}(u, v) + \mathrm{lso}(v, u)$ (see Corollary 84). Then further classify all pairs (u, v) into two groups based on $\mathrm{lso}(u, v) + \mathrm{lso}(v, u)$. If $\mathrm{lso}(u, v) + \mathrm{lso}(v, u)$ is 'small', then we can easily count the number of such pairs merely by using the number of unbordered words (see Lemma 82). If $\mathrm{lso}(u, v) + \mathrm{lso}(v, u)$ is 'large', then the pair (u, v) has a certain structure we can exploit to count them (see Lemma 83).

Lemma 81. *Let $n \geq 1$. Let $u, v \in \Sigma_k^n$. Let w be a non-empty word that is both a proper suffix of u and a proper prefix of v. Then $w = \mathrm{so}(u, v)$ iff w is unbordered.*

Proof. We prove an equivalent proposition, namely that $w \neq \mathrm{so}(u, v)$ iff w is bordered.

$w \neq \mathrm{so}(u, v) \iff$ There exists a non-empty word x such that $|x| < |w|$ and $x = \mathrm{so}(u, v)$. $\iff x = \mathrm{so}(w, w) \iff w = xs = tx$ for some $s, t \in \Sigma_k^+ \iff w$ is bordered. □

Table 8.4: Some values of $M_2(n)$, $R_2(n)$, and $U_2(n)$ for n where $1 \leq n \leq 15$.

n	$M_2(n)$	$R_2(n)$	$U_2(n)$
1	0	0	4
2	4	4	4
3	26	14	10
4	124	52	28
5	524	204	92
6	2154	806	330
7	8706	3214	1250
8	34996	12844	4852
9	140290	51366	19122
10	561724	205492	75868
11	2247892	822108	302196
12	8993414	3288858	1206086
13	35976928	13156624	4818688
14	143913546	52629590	19262730
15	575664422	210525818	77025766

Lemma 82. *Let $n \geq 1$. Let u and v be length-n words. Let $i = \mathrm{lso}(u, v)$, and $j = \mathrm{lso}(v, u)$. Then $i + j \leq n$ iff $u = xsy$ and $v = ytx$ for some words $s, t \in \Sigma_k^*, x \in \mathcal{U}_j^k$, and $y \in \mathcal{U}_i^k$.*

Proof.

$\Longrightarrow$: Let $y = \mathrm{so}(u, v)$ and $x = \mathrm{so}(v, u)$. Let $j = \mathrm{lso}(u, v)$ and $i = \mathrm{lso}(v, u)$. Suppose $i + j \leq n$. By definition there exist words $w, z, \alpha, \beta \in \Sigma_k^*$ such that $u = wy, v = yz, v = \alpha x$, and $u = x\beta$. But since $i + j = |x| + |y| \leq n$, we have that x and y do not overlap. Thus there exist words $s, t \in \Sigma_k^{n-i-j}$ such that $u = xsy$ and $v = ytx$. By Lemma 81, we have that x and y must be unbordered. Therefore $x \in \mathcal{U}_j^k$ and $y \in \mathcal{U}_i^k$

$\Longleftarrow$: Follows from the definition of u and v. □

Lemma 83. *Let $n \geq 1$. Let u and v be length-n words. Let $i = \mathrm{lso}(u, v)$, and $j = \mathrm{lso}(v, u)$. Then $i + j > n$ iff*

a) $n + 1 \leq i + j \leq \frac{4}{3}n$, and

b) there exist distinct words $x, y \in \Sigma_k^{i+j-n}$, and $s, t \in \Sigma_k^*$ such that u is of the form $xsytx$ and v is of the form $ytxsy$ where (x, y) is mutually unbordered, and both xsy and ytx are unbordered with $\mathrm{so}(u, v) = ytx$ and $\mathrm{so}(v, u) = xsy$.

Proof.

$\Longrightarrow$: Since $i+j > n$ we have that $(\text{so}(u,v), \text{so}(v,u))$ has a right-border and a left-border. Let x be the length-$(i+j-n)$ suffix of $\text{so}(u,v)$. We can now write $u = r\alpha x = xw\beta$ and $v = \alpha xw$ for some $r, w, \alpha, \beta \in \Sigma_k^*$ where $|\alpha x| = i$ and $|xw| = j$. Clearly x is a prefix and suffix of u. Let y be the length-$(i+j-n)$ suffix of $\text{so}(v,u)$. By a similar argument as above, one can show that y is a prefix and suffix of v. If (x,y) has a right-border or a left-border, then (u,v) has a right-border of length $< i$ or a left-border of length $< j$. So (x,y) must be mutually unbordered and u must be of the form $xsytx$ and v must be of the form $ytxsy$ for some $s, t \in \Sigma_k^*$ where $|ytx| = i$ and $|xsy| = j$. Since (x,y) is mutually unbordered, the words u and v can only be of this form if $2|x| + |y| = 2|y| + |x| \leq n$. Since $|x| = |y| = (i+j-n)$, we have that $3(i+j-n) \leq n \implies i+j \leq \frac{4}{3}n$. Now by Lemma 81, we have that ytx and xsy are unbordered. Since both xsy and ytx are unbordered and $|x| = |y|$, we have that y and x must be distinct.

$\Longleftarrow$: We have that $i+j > n$ by assumption. □

Perhaps the most peculiar and interesting aspect to Lemma 83 is the fact that for length-n words u and v, the sum of $\text{lso}(u,v)$ and $\text{lso}(v,u)$ is bounded by a number between n and $2n$. This fact is outlined in Corollary 84.

Corollary 84. *Let $n \geq 1$. Let u and v be length-n words. Then* $\text{lso}(u,v) + \text{lso}(v,u) \leq \frac{4}{3}n$.

In Example 85 we illustrate pairs of words (u,v) that attain the bound $\frac{4}{3}n$ bound.

Example 85. The following three pairs of words illustrate the upper bound $\lfloor \frac{4}{3}n \rfloor$ from Lemma 83 and Corollary 84. We give examples for all lengths of words by giving examples for each equivalence class modulo 3.

For $n = 3m$, we have

$$(u,v) = (0^m 1^m 0^m, 1^m 0^m 1^m).$$

For $n = 3m+1$, we have

$$(u,v) = (0^m 1^{m+1} 0^m, 1^{m+1} 0^m 1^m).$$

For $n = 3m+2$, we have

$$(u,v) = (0^m 1^{m+2} 0^m, 1^{m+2} 0^m 1^m).$$

Lemma 83 shows that pairs of length-n non-empty words (u, v) where $\text{lso}(u, v)+\text{lso}(v, u)$ is 'large' ($> n$) exhibit a particular structure. Namely $\text{so}(u, v)$ is unbordered and $\text{so}(u, v)$ begins and ends with a mutually unbordered pair of words. The same is true for $\text{so}(v, u)$ as well. So we need an expression for the number of unbordered words whose prefix and suffix of a certain length form a pair of mutually unbordered words.

Let $t \leq n$ be a positive integer. Let u and v be length-t words such that (u, v) is mutually unbordered. Let $G_{u,v}(n)$ denote the number of length-n unbordered words that have u as a prefix, v as a suffix (and vice versa).

Lemma 86. *Let $n \geq t \geq 1$. Let u and v be length-t words such that (u, v) is mutually unbordered and $u \neq v$. Then the number of unbordered words that have u as a prefix and v as a suffix is*

$$G_{u,v}(n) = \begin{cases} 0, & \text{if } n < 2t; \\ k^{n-2t} - \sum_{i=2t}^{\lfloor n/2 \rfloor} G_{u,v}(i)k^{n-2i}, & \text{if } n \geq 2t. \end{cases}$$

Proof. If $n < 2t$ then $G_{u,v}(n) = 0$, since (u, v) is mutually unbordered. Suppose $n \geq 2t$. Then the number of unbordered words of length n having u as a prefix and v as a suffix is equal to the number of bordered words that contain u as a prefix and v as a suffix, subtracted from the total number of words that contain u as a prefix and v as a suffix. Let w be a word of length n such that u is a prefix of w and v is a suffix of w. Then w is bordered if and only if its shortest border is of length j where $2t \leq j \leq \lfloor n/2 \rfloor$. This is because (u, v) is mutually unbordered and words that have a border of length $> n/2$ must also have a border of length $\leq n/2$. Also observe that the shortest border of w must itself be unbordered and have u as a prefix and v as a suffix. So the total number of words of the form w as described above is $\sum_{i=2t}^{\lfloor n/2 \rfloor} G_{u,v}(i)k^{n-2i}$. Therefore, for $n \geq 2t$ we have

$$G_{u,v}(n) = k^{n-2t} - \sum_{i=2t}^{\lfloor n/2 \rfloor} G_{u,v}(i)k^{n-2i}.$$

□

From Lemma 86 we see that $G_{u,v}$ is independent of u and v, but dependent on the length of $|u| = |v|$. Therefore, for u, v words of length t, let $G_{u,v}(n) = G_t(n)$.

Finally, we are ready to present recurrences for $M_k(n), R_k(n)$, and $U_k(n)$.

Theorem 87. *The number $M_k(n)$ of mutually bordered pairs of words of equal length satisfies*

$$M_k(n) = \sum_{i=1}^{n-1} \sum_{j=1}^{n-i} u_i u_j k^{2n-2(i+j)} + \sum_{i=1}^{\lfloor n/3 \rfloor} (U_k(i) - u_i) \sum_{j=2i}^{n-i} G_i(j) G_i(n-j+i), \tag{8.1}$$

where $M_k(1) = 0$. Additionally we have

$$R_k(n) = \left(\sum_{i=1}^{n-1} k^{2n-2i} u_i \right) - M_k(n), \tag{8.2}$$

and

$$U_k(n) + 2R_k(n) + M_k(n) = k^{2n}. \tag{8.3}$$

Proof. Let $n \geq 1$, and let u and v be words of length n.

Proof of Eq. (8.3): Let (u, v) be a pair of length-n words. Exactly one of the following must be true about (u, v):

(a) (u, v) has a right-border and a left-border (i.e., (u, v) is mutually bordered),

(b) (u, v) has a right-border but not a left-border (i.e., (u, v) is right-bordered),

(c) (u, v) does not have a right-border but has a left-border (i.e., (u, v) is left-bordered),

(d) (u, v) does not have a right-border or a left-border (i.e., (u, v) is mutually unbordered).

Clearly the number of right-bordered pairs of words is the same as the number of pairs of left-bordered pairs of words. From these facts we conclude that

$$U_k(n) + 2R_k(n) + M_k(n) = k^{2n}.$$

Proof of Eq. (8.2): The number of right-bordered pairs of words is equal to the total number of pairs (u, v) that have a right-border subtracted from the total number of mutually bordered pairs of words. So

$$R_k(n) = \left(\sum_{i=1}^{n-1} k^{2n-2i} u_i \right) - M_k(n).$$

Proof of Eq. (8.1): Clearly $M(1) = 0$ since words of length 1 cannot have left-borders or right-borders. Let $i = \mathrm{lso}(u, v)$ and $j = \mathrm{lso}(v, u)$. By Lemma 82 we have that $i + j \leq n$ iff $u = xsy$ and $v = ytx$ for some words $s, t \in \Sigma_k^*, x \in \mathcal{U}_j^k$, and $y \in \mathcal{U}_i^k$. We can count the number of pairs of such words using the number of unbordered words and summing over all possible lengths of s, y, x, and y. We get

$$\sum_{i=1}^{n-1} \sum_{j=1}^{n-i} u_i u_j k^{2n-2(i+j)}.$$

By Lemma 83 we have that $i + j > n$ iff $n + 1 \leq i + j \leq \frac{4}{3}n$ and there exist words $x, y \in \Sigma_k^{i+j-n}$, and $s, t \in \Sigma_k^*$ such that u is of the form $xsytx$ and v is of the form $ytxsy$ where (x, y) is mutually unbordered, both xsy and ytx are unbordered with $|ytx| = i$ and $|xsy| = j$, and $x \neq y$. The fact that $n + 1 \leq i + j \leq \frac{4}{3}n$ and i, j, n are integers implies that $1 \leq |x| = |y| = (i + j - n) \leq \lfloor n/3 \rfloor$. Let $p = (i + j - n)$. Since (x, y) is mutually unbordered and both xsy and ytx are unbordered, we have $x \neq y$. Suppose that we in fact have $w = x = y$ for some $w \in \Sigma_k^p$. The only such w must be unbordered, since (x, y) is unbordered. Therefore, the number of mutually unbordered pairs (x, y) with $x \neq y$ is $U_k(p) - u_p$.

By Lemma 86 we know that for (x, y) fixed, the number of unbordered words of the form ytx with (x, y) mutually unbordered and $x \neq y$ is $G_{x,y}(i) = G_p(i)$. Similarly the number of unbordered words of the form xsy is $G_{x,y}(n - i + p) = G_p(n - i + p)$. We also have that $i \geq 2p$ and $i \leq n - p$ since (x, y) is mutually unbordered. For (x, y) fixed, the total number of pairs of words of the form (xsy, ytx) is

$$\sum_{l=2p}^{n-p} G_p(l) G_p(n - l + p).$$

So the number of words of the form $xsytx$ and $ytxsy$ as described above is equal to the number of pairs of words (xsy, ytx) with (x, y) mutually unbordered, xsy and ytx unbordered, and $x \neq y$. Since $1 \leq p = (i + j - n) \leq \lfloor n/3 \rfloor$ we have that this is equal to

$$\sum_{p=1}^{\lfloor n/3 \rfloor} (U_k(p) - u_p) \sum_{l=2p}^{n-p} G_p(l) G_p(n - l + p).$$

Putting it all together, we have

$$M_k(n) = \left(\sum_{i=1}^{n-1} \sum_{j=1}^{n-i} u_i u_j k^{2n-2(i+j)} \right) + \sum_{i=1}^{\lfloor n/3 \rfloor} (U_k(i) - u_i) \sum_{j=2i}^{n-i} G_i(j) G_i(n - j + i).$$

□

8.3 Limiting values

In this section we show that the limit $L_k = \lim_{n\to\infty} M_k(n)/k^{2n}$ exists.

Theorem 88. *The following limit exists:*

$$L_k = \lim_{n\to\infty} \frac{M_k(n)}{k^{2n}}.$$

Furthermore, we have that

$$\left(\sum_{i=1}^{n} u_i k^{-2i}\right)^2 \leq L_k \leq \left(\left(\sum_{i=1}^{n-1} u_i k^{-2i}\right) + \frac{k^{-n}}{k-1}\right)^2.$$

Proof. From the recurrence for $M_k(n)$, we see that there are two main terms. The first term is

$$\sum_{i=1}^{n-1}\sum_{j=1}^{n-i} u_i u_j k^{2n-2(i+j)},$$

which counts all pairs of words (u, v) where $\mathrm{lso}(u, v) + \mathrm{lso}(v, u)$ is 'small'. The second term is

$$\sum_{i=1}^{\lfloor n/3\rfloor} (U_k(i) - u_i) \sum_{j=2i}^{n-i} G_i(j) G_i(n-j+i),$$

which counts all pairs of words (u, v) where $\mathrm{lso}(u, v) + \mathrm{lso}(v, u)$ is 'large'. Now from Lemma 83, we know that the only pairs (u, v) where $\mathrm{lso}(u, v) + \mathrm{lso}(v, u)$ is 'large' are pairs of words of the form $(xsytx, ytxsy)$. There are at most k^n choices for $xsytx$, and $ytxsy$ can be recreated from $xsytx$ by knowing the starting positions of s, y, and t. Therefore there must be $o(k^{2n})$ pairs of such words. So in the limit L_k, this term goes to 0, and thus

$$L_k = \lim_{n\to\infty} \sum_{i=1}^{n-1}\sum_{j=1}^{n-i} u_i u_j k^{-2(i+j)}.$$

Consider the infinite double series

$$L'_k = \lim_{n\to\infty} \sum_{i=1}^{n-1}\sum_{j=1}^{n-1} u_i u_j k^{-2(i+j)}.$$

We can factor out $u_i k^{-2i}$ out of the nested series and split up the nested sum to get

$$\begin{aligned} L'_k &= \lim_{n\to\infty} \sum_{i=1}^{n-1} u_i k^{-2i} \sum_{j=1}^{n-1} u_j k^{-2j} \\ &= \lim_{n\to\infty} \sum_{i=1}^{n-1} u_i k^{-2i} \left(\sum_{j=1}^{n-i} u_j k^{-2j} + \sum_{j=n-i+1}^{n-1} u_j k^{-2j} \right). \end{aligned}$$

For each i, we have $\lim_{n\to\infty} \sum_{j=n-i+1}^{n-1} u_j k^{-2j} = 0$, and thus, $L_k = L'_k$.

So using the fact that $L_k = L'_k$, we have

$$L_k = \left(\sum_{i=1}^{\infty} u_i k^{-2i} \right)^2.$$

Thus the limit L_k exists if and only if the series $\sum_{i=1}^{\infty} u_i k^{-2i}$ converges. Since $u_i \leq k^i$ we have that $u_i k^{-2i} \leq k^{-i}$. Therefore by direct comparison we have that $\sum_{i=1}^{\infty} u_i k^{-2i}$ converges, and so the limit L_k exists. Since $u_i k^{-2i} \leq k^{-i}$ we have that

$$\sum_{i=m}^{\infty} u_i k^{-2i} \leq \sum_{i=m}^{\infty} k^{-i}.$$

Using this we get the following inequalities,

$$\begin{aligned} \sum_{i=1}^{n} u_i k^{-2i} \leq \sum_{i=1}^{\infty} u_i k^{-2i} &\leq \left(\sum_{i=1}^{n-1} u_i k^{-2i} \right) + \sum_{i=n}^{\infty} k^{-i} \\ &= \left(\sum_{i=1}^{n-1} u_i k^{-2i} \right) + \frac{k^{-n}}{k-1}. \end{aligned}$$

Now we have bounds for our limit,

$$\begin{aligned} \left(\sum_{i=1}^{n} u_i k^{-2i} \right)^2 \leq L_k &= \left(\sum_{i=1}^{\infty} u_i k^{-2i} \right)^2 \\ &\leq \left(\left(\sum_{i=1}^{n-1} u_i k^{-2i} \right) + \frac{k^{-n}}{k-1} \right)^2. \end{aligned}$$

□

Corollary 89. *The following limit exists:*

$$\lim_{n\to\infty} \frac{R_k(n)}{k^{2n}}.$$

Corollary 90. *The following limit exists:*

$$\lim_{n\to\infty} \frac{U_k(n)}{k^{2n}}.$$

Table 8.5 shows the behaviour of the functions $M_k(n)$, $R_k(n)$, and $U_k(n)$ as k increases. Interestingly, there are more mutually bordered pairs than not when $k = 2$, but when k increases the number of mutually unbordered pairs of words dominates.

k	$\lim_{n\to\infty} M_k(n)/k^{2n}$	$\lim_{n\to\infty} R_k(n)/k^{2n}$	$\lim_{n\to\infty} U_k(n)/k^{2n}$
2	0.536	0.196	0.072
3	0.196	0.247	0.310
4	0.098	0.215	0.473
5	0.058	0.182	0.578
⋮	⋮	⋮	⋮
10	0.012	0.098	0.792
⋮	⋮	⋮	⋮
100	0.000	0.010	0.980

Table 8.5: Limits of recurrences as k increases.

8.4 Expected shortest right-border

In this section we compute expected value of $\mathrm{lso}(u, v)$ and $\mathrm{lso}(v, u)$ for length-n words u and v. Additionally, we show that the expected value tends to a constant.

Let $S_k(i, n)$ denote the number of pairs of length-n words (u, v) over a k-letter alphabet such that $\mathrm{lso}(u, v) = i$.

Proposition 91. *Let $n, k, i \geq 1$. Then $S_k(i, n) = u_i k^{2(n-i)}$.*

Proof. Follows directly from Lemma 81. □

Theorem 92. *Let $n, k \geq 1$. Let u and v be length-n words over a k-letter alphabet. Then the expected value of* $\mathrm{lso}(u, v)$ *is* $O(1)$.

Proof.

$$\begin{aligned}\lim_{n\to\infty} \sum_{i=0}^{n} i \cdot \Pr[X = i] &= \lim_{n\to\infty} \frac{1}{k^{2n}} \sum_{i=1}^{n-1} i \cdot S_k(i, n) \\ &= \sum_{i=1}^{\infty} i \cdot u_i k^{-2i}.\end{aligned}$$

Since $u_i \leq k^i$, we have that $i \cdot u_i k^{-2i} \leq i \cdot k^{-i}$. Therefore by direct comparison, the series $\sum_{i=1}^{\infty} i \cdot u_i k^{-2i}$ converges.

Since $i \cdot u_i k^{-2i} \leq i \cdot k^{-i}$ we have that $\sum_{i=m}^{\infty} i \cdot u_i k^{-2i} \leq \sum_{i=m}^{\infty} i \cdot k^{-i}$. Using this we get the bounds

$$\begin{aligned}\sum_{i=1}^{n} i \cdot u_i k^{-2i} \leq \sum_{i=1}^{\infty} i \cdot u_i k^{-2i} &\leq \left(\sum_{i=1}^{n} i \cdot u_i k^{-2i} \right) + \sum_{i=n+1}^{\infty} i \cdot k^{-i} \\ &= \left(\sum_{i=1}^{n} i \cdot u_i k^{-2i} \right) + k^{-n} \frac{k(n+1) - n}{(1-k)^2}.\end{aligned}$$

□

Table 8.6 shows the behaviour of the expected shortest right-border/left-border as k increases. Interestingly, $k = 2$ is the only value of k for which the expected length of the shortest left-border and right-border is greater than 1. For all other $k > 2$ we have that the expected shortest left-border and right-border are less than one symbol in length.

k	$\sum_{i=1}^{\infty} i \cdot u_i k^{-2i}$
2	1.156
3	0.605
4	0.395
5	0.290
⋮	⋮
10	0.121
⋮	⋮
100	0.010

Table 8.6: Asymptotic expected value of $\mathrm{lso}(u, v)$ and $\mathrm{lso}(v, u)$.

Chapter 9

Conclusions and open problems

In this thesis we have studied many enumeration problems and properties of variations and generalizations of bordered words.

We started by introducing the field of combinatorics on words and briefly mentioning the literature on borders relevant to this thesis in Chapter 1.

Then in Chapter 2 we gave necessary definitions and went more in depth on the relevant literature on borders.

In Chapter 3, using software that implements a decision procedure for k-automatic sequences, we completed the characterization due to Harju and Nowotka [56, 60] of binary words with the maximum number $\text{mnuc}_2(n)$ of unbordered conjugates. We also showed that for any number, up to the maximum $\text{mnuc}_2(n)$, there exists a word with that number of unbordered conjugates.

In Chapter 4 we characterized and counted all pairs of words x and y almost commute (i.e., xy and yx differ in exactly two positions). We also characterized and counted pairs of words x and y that anti-commute (i.e., xy and yx differ in all positions). When counting pairs of words that almost commute, we were able to count all almost-commuting pairs of words (x, y) with $|xy| = n$. This was because of Lemma 46, an analogue of the Fine-Wilf theorem that characterizes all pairs x, y that can be broken up such that $xy = x'y'$ with $x \neq x'$ where x' and y' almost commute. We did not see a way to prove something similar for anti-commuting pairs of words. So we pose the following open problems:

- Characterize and count all pairs of words x and y such that $\text{ham}(xy, yx) = m$.
- Find a recurrence to count the number of anti-commuting pairs of words (x, y) with $|xy| = n$.

In Chapter 5 we characterize and count all words with a unique border. We also show that the probability $P_{n,k}$ that a randomly chosen length-n word has a unique border tends to a constant. We pose the following open problems and questions:

- Find good bounds on $P_{n,k}$.
- Given a word w with a unique border, what is the expected length of this border?

In Chapter 6 we characterized and counted all length-n words that have a width-t largest block palindrome. We also showed that the expected width E_k of a word's largest block palindrome tends to a constant. Finally we defined the smallest block palindrome and proved bounds for the width of the smallest block palindrome of a word. We pose the following open problems and questions:

- Find a tight upper bound on the asymptotic expected bound E_k of the largest block palindrome of a word.
- How many length-n words have a smallest block palindrome, in the non-overlapping sense, of width t?
- What is the expected width of a word's smallest block palindrome?

In Chapter 7 presented the two main results of this thesis. We improved existing bounds for $C_k(n)$ and $P_k(n)$. We showed that $C_k(n) \in \Theta(\frac{k^n}{n})$. In other words, we showed that $C_k(n)$ can be bounded above and below by a constant times k^n/n for n sufficiently large. We ask the following questions:

- Can we do better than this? Does the limit

$$\lim_{n\to\infty} \frac{C_k(n)}{k^n/n}$$

  exist? If it does exist, what does the limit evaluate to? Can one find good bounds on the limit?

We also gave a family of upper and lower bounds for $P_k(n)$ in Chapter 7. But for every $j \geq 0$, the upper and lower bounds on $P_k(n)$ are asymptotically separated by a factor of $1/\log_k^{\circ j}(n)$. We ask the following questions:

- Does there exist a $g(n)$ such that $P_k(n) \in \Theta(\frac{k^n}{g(n)})$? If such a function $g(n)$ exists, then does the limit
$$\lim_{n\to\infty} \frac{P_k(n)}{k^n/g(n)}$$
exist?

In Chapter 8 we gave recurrences and asymptotic results for the number of pairs of mutually unbordered and bordered words. But we only proved these results for equal-length pairs of words. Additionally, the way we proved these recurrences was by looking at the shortest right-border/left-border of pairs of words. We pose the following open questions and problems:

- Find recurrences for $M_k(n)$, $R_k(n)$, and $U_k(n)$ that are not coupled.
- How many pairs of length-n words (u, v) have a largest right-border/left-border of length i?
- What is the expected length of the longest right-border/left-border of a pair of words?
- Find recurrences for $M_k(m, n)$, $R_k(m, n)$, and $U_k(m, n)$.
- Generalize to arbitrary tuples or sets of size ≥ 3. Two obvious generalizations possible:
    1. All consecutive words in a tuple are either mutually unbordered, or mutually bordered.
    2. All pairs of words in a set are either mutually unbordered (i.e., cross-bifix-free sets of words), or mutually bordered.
- The term *Gray code* is used to describe an exhaustive listing of a set of combinatorial objects where successive terms differ by some small, well-defined amount. Gray codes are named after Frank Gray, who discovered a simple method of listing all 2^n binary words where successive words differ in exactly one position. Gray codes for cross-bifix-free sets already exist [15, 14]. Can one generate a Gray code for mutually unbordered and bordered pairs of length-n words where successive terms differ by a small amount if their Hamming distance is bounded by some constant $C \geq 1$? How small can C get?

- The concept of mutual borderedness and unborderedness can be extended to two dimensions [9, 11] where words are two-dimensional matrices with entries taken from a finite alphabet. In analogy with the study of mutually bordered and unbordered pairs of words in this thesis, can one do the same for the set of all mutually bordered and unbordered pairs of $p \times q$-matrices?

This chapter concludes the work.

References

[1] The 2015 British Informatics Olympiad (Round 1 Question 1). https://olympiad.org.uk/2015/index.html.

[2] J.-P. Allouche and J. Shallit. *Automatic Sequences: Theory, Applications, Generalizations*. Cambridge University Press, 2003.

[3] J.-P. Allouche and J. O. Shallit. The ubiquitous Prouhet-Thue-Morse sequence. In C. Ding, T. Helleseth, and H. Niederreiter, editors, *Sequences and Their Applications, Proceedings of SETA '98*, Discrete Math. & Theoret. Comput. Sci., pages 1–16. Springer-Verlag, 1999.

[4] M. Anselmo, D. Giammarresi, and M. Madonia. Unbordered pictures: Properties and construction. In Andreas Maletti, editor, *Algebraic Informatics*, volume 9270 of *Lecture Notes in Computer Science*, pages 45–57, Cham, 2015. Springer International Publishing.

[5] R. Assous and M. Pouzet. Une caractérisation des mots périodiques. *Discrete Math.*, 25(1):1–5, 1979.

[6] G. Badkobeh, G. Fici, and Z. Lipták. On the number of closed factors in a word. In Adrian-Horia Dediu, Enrico Formenti, Carlos Martín-Vide, and Bianca Truthe, editors, *Language and Automata Theory and Applications*, volume 8977 of *Lecture Notes in Computer Science*, pages 381–390, Cham, 2015. Springer International Publishing.

[7] D. Bajić and T. Loncar-Turukalo. A simple suboptimal construction of cross-bifix-free codes. *Cryptography and Communications*, 6:27–37, 2014.

[8] D. Bajić and J. Stojanović. Distributed sequences and search process. In *2004 IEEE International Conference on Communications*, volume 1, pages 514–518, 2004.

[9] E. Barcucci, A. Bernini, S. Bilotta, and R. Pinzani. Cross-bifix-free sets in two dimensions. *Theoret. Comput. Sci.*, 664:29–38, 2017.

[10] E. Barcucci, A. Bernini, S. Bilotta, and R. Pinzani. Non-overlapping matrices. *Theoret. Comput. Sci.*, 658:36–45, 2017.

[11] E. Barcucci, A. Bernini, S. Bilotta, and R. Pinzani. A 2D non-overlapping code over a q-ary alphabet. *Cryptography and Communications*, 10(4):667–683, July 2018.

[12] E. Barcucci, A. Bernini, and R. Pinzani. A strong non-overlapping Dyck code. In Nelma Moreira and Rogério Reis, editors, *Developments in Language Theory*, volume 12811 of *Lecture Notes in Computer Science*, pages 43–53, Cham, 2021. Springer International Publishing.

[13] M.-P. Béal and M. Crochemore. Checking whether a word is Hamming-isometric in linear time. *Theoret. Comput. Sci.*, 933:55–59, 2022.

[14] A. Bernini, S. Bilotta, R. Pinzani, A. Sabri, and V. Vajnovszki. Prefix partitioned gray codes for particular cross-bifix-free sets. *Cryptography and Communications*, 6(4):359–369, Dec 2014.

[15] A. Bernini, S. Bilotta, R. Pinzani, and V. Vajnovszki. A Gray code for cross-bifix-free sets. *Mathematical Structures in Computer Science*, 27(2):184–196, 2017.

[16] J. Berstel and D. Perrin. *Theory of codes*, volume 117 of *Pure and Applied Mathematics*. Academic Press Inc., 1985.

[17] S. Bilotta. Variable-length non-overlapping codes. *IEEE Trans. Inform. Theory*, 63(10):6530–6537, 2017.

[18] S. Bilotta, E. Pergola, and R. Pinzani. A new approach to cross-bifix-free sets. *IEEE Trans. Inform. Theory*, 58(6):4058–4063, June 2012.

[19] S. R. Blackburn. Non-overlapping codes. *IEEE Trans. Inform. Theory*, 61(9):4890–4894, 2015.

[20] G. Blom. Overalapping binary sequences (problem 94-20). *SIAM Review*, 37(4):619–620, 1995.

[21] R. S. Boyer and J. S. Moore. A fast string searching algorithm. *Commun. ACM*, 20(10):762–772, October 1977.

[22] V. Bruyère, G. Hansel, C. Michaux, and R. Villemaire. Logic and p-recognizable sets of integers. *Bull. Belgian Math. Soc.*, 1:191–238, 1994. Corrigendum, Bull. Belg. Math. Soc. **1**, 577 (1994).

[23] M. Bucci, A. de Luca, and A. De Luca. Rich and periodic-like words. In Volker Diekert and Dirk Nowotka, editors, *Developments in Language Theory*, volume 5583 of *Lecture Notes in Computer Science*, pages 145–155, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[24] M. Bucci, A. De Luca, and G. Fici. Enumeration and structure of trapezoidal words. *Theoret. Comput. Sci.*, 468:12–22, 2013.

[25] A. Carpi and A. de Luca. Periodic-like words, periodicity, and boxes. *Acta Informatica*, 37(8):597–618, May 2001.

[26] Y. M. Chee, H. M. Kiah, P. Purkayastha, and C. Wang. Cross-bifix-free codes within a constant factor of optimality. *IEEE Trans. Inform. Theory*, 59(7):4668–4674, July 2013.

[27] T. Clokie, D. Gabric, and J. Shallit. Circularly squarefree words and unbordered conjugates: A new approach. In *Combinatorics on Words*, volume 11682 of *Lecture Notes in Computer Science*, pages 133–144. Springer, Cham, 2019.

[28] P. H. Cording, T. Gagie, M. B. T. Knudsen, and T. Kociumaka. Maximal unbordered factors of random strings. *Theoret. Comput. Sci.*, 852:78–83, 2021.

[29] J. C. Costa. Biinfinite words with maximal recurrent unbordered factors. *Theoret. Comput. Sci.*, 290:2053–2061, 2003.

[30] A. J. de Lind van Wijngaarden and T. J. Willink. Frame synchronization using distributed sequences. *IEEE Trans. Commun.*, 48(12):2127–2138, 2000.

[31] A. De Luca and G. Fici. Open and closed prefixes of Sturmian words. In Juhani Karhumäki, Arto Lepistö, and Luca Zamboni, editors, *Combinatorics on Words*, volume 8079 of *Lecture Notes in Computer Science*, pages 132–142, Berlin, Heidelberg, 2013. Springer.

[32] A. de Luca and F. Mignosi. Some combinatorial properties of Sturmian words. *Theoret. Comput. Sci.*, 136(2):361–385, 1994.

[33] J.-P. Duval. Une caractérisation de la période d'un mot fini par la longueur de ses facteurs primaires. *C. R. Acad. Sci. Paris*, 290:A359–A361, 1980.

[34] J.-P. Duval. Relationship between the period of a finite word and the length of its unbordered segments. *Discrete Math.*, 40:31–44, 1982.

[35] J.-P. Duval, T. Harju, and D. Nowotka. Unbordered factors and Lyndon words. *Discrete Math.*, 308:2261–2264, 2008.

[36] A. Ehrenfeucht and D. M. Silberger. Periodicity and unbordered segments of words. *Discrete Math.*, 26:101–109, 1979.

[37] G. Fici. A classification of trapezoidal words. In P. Ambrož, Š. Holub, and Z. Masáková, editors, *8th International Conference on Words, WORDS 2011*, volume 63 of *Electronic Proceedings in Theoretical Computer Science*, pages 129–137. 2011.

[38] G. Fici. Open and closed words. *Bulletin of EATCS*, 3(123):140–149, 2017.

[39] N. J. Fine and H. S. Wilf. Uniqueness theorems for periodic functions. *Proc. Amer. Math. Soc.*, 16(1):109–114, 1965.

[40] M. Forsyth, A. Jayakumar, J. Peltomäki, and J. Shallit. Remarks on privileged words. *Internat. J. Found. Comp. Sci.*, 27(04):431–442, 2016.

[41] A. E. Frid, S. Puzynina, and L. Q. Zamboni. On palindromic factorization of words. *Adv. in Appl. Math.*, 50(5):737–748, 2013.

[42] D. Gabric. Asymptotic bounds for the number of closed and privileged words, 2022. arXiv:2206.14273. Link: https://arxiv.org/abs/2206.14273.

[43] D. Gabric. Mutual borders and overlaps. *IEEE Trans. Inform. Theory*, 68(10):6888–6893, 2022.

[44] D. Gabric. Words that almost commute. *Discrete Math.*, 345(112898):1–8, 2022.

[45] D. Gabric, N. Rampersad, and J. Shallit. An inequality for the number of periods in a word. *Internat. J. Found. Comp. Sci.*, 32(05):597–614, 2021.

[46] D. Gabric and J. Shallit. Borders, palindrome prefixes, and square prefixes. *Inform. Process. Lett.*, 165:106027, 2021.

[47] E. Gilbert. Synchronization of binary messages. *IRE Trans. Info. Theory*, 6(4):470–477, 1960.

[48] A. Glen, J. Justin, S. Widmer, and L. Q. Zamboni. Palindromic richness. *European J. Combinatorics*, 30(2):510–531, 2009.

[49] K. Goto, I. Tomohiro, H. Bannai, and S. Inenaga. Block palindromes: A new generalization of palindromes. In T. Gagie, A. Moffat, G. Navarro, and E. Cuadros-Vargas, editors, *String Processing and Information Retrieval*, volume 11147 of *Lecture Notes in Computer Science*, pages 183–190, Cham, 2018. Springer International Publishing.

[50] L. J. Guibas and A. M. Odlyzko. Maximal prefix-synchronized codes. *SIAM J. Appl. Math.*, 35(2):401–418, 1978.

[51] L. J. Guibas and A. M. Odlyzko. Periods in strings. *J. Combin. Theory Ser. A*, 30(1):19–42, 1981.

[52] L. J. Guibas and A. M. Odlyzko. String overlaps, pattern matching, and nontransitive games. *J. Combin. Theory Ser. A*, 30(2):183–208, 1981.

[53] R. W. Hamming. Error detecting and error correcting codes. *Bell System Tech. J.*, 29(2):147–160, 1950.

[54] Y.-S. Han and K. Salomaa. Overlap-free languages and solid codes. *Internat. J. Found. Comp. Sci.*, 22(05):1197–1209, 2011.

[55] Y.-S. Han and D. Wood. Overlap-free regular languages. In D. Z. Chen and D. T. Lee, editors, *Computing and Combinatorics*, volume 4112 of *Lecture Notes in Computer Science*, pages 469–478, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[56] T. Harju and D. Nowotka. Border correlation of binary words. *J. Combin. Theory Ser. A*, 108:331–341, 2004.

[57] T. Harju and D. Nowotka. Periodicity and unbordered words. In Volker Diekert and Michel Habib, editors, *STACS 2004*, volume 2996 of *Lecture Notes in Computer Science*, pages 294–304, Berlin, Heidelberg, 2004. Springer.

[58] T. Harju and D. Nowotka. Counting bordered and primitive words with a fixed weight. *Theoret. Comput. Sci.*, 340(2):273–279, 2005.

[59] T. Harju and D. Nowotka. Periodicity and unbordered words: a proof of the extended Duval conjecture. *J. Assoc. Comput. Mach.*, 54:1–20, 2007.

[60] T. Harju and D. Nowotka. Bordered conjugates of words over large alphabets. *Electronic J. Combinatorics*, 15, 2008. #N41.

[61] Š. Holub. A proof of the extended Duval's conjecture. *Theoret. Comput. Sci.*, 339:61–67, 2005.

[62] Š. Holub and M. Müller. Fully bordered words. *Theoret. Comput. Sci.*, 684:53–58, 2017.

[63] Š. Holub and D. Nowotka. On the relation between periodicity and unbordered factors of finite words. *Internat. J. Found. Comp. Sci.*, 21:633–645, 2010.

[64] Š. Holub and D. Nowotka. The Ehrenfeucht–Silberger problem. *J. Combin. Theory Ser. A*, 119(3):668–682, 2012.

[65] Š. Holub and J. Shallit. Periods and borders of random words. In Nicolas Ollinger and Heribert Vollmer, editors, *33rd Symposium on Theoretical Aspects of Computer Science (STACS 2016)*, volume 47 of *Leibniz International Proceedings in Informatics*, pages 44:1–44:10. Schloss Dagstuhl—Leibniz Center for Informatics, 2016.

[66] M. Jahannia, M. Mohammad-Noori, N. Rampersad, and M. Stipulanti. Closed Ziv–Lempel factorization of the m-bonacci words. *Theoret. Comput. Sci.*, 918:32–47, 2022.

[67] J. Kellendonk, D. Lenz, and J. Savinien. A characterization of subshifts with bounded powers. *Discrete Math.*, 313(24):2881–2894, 2013.

[68] S. Klavžar and S. Shpectorov. Asymptotic number of isometric generalized Fibonacci cubes. *European J. Combinatorics*, 33(2):220–226, February 2012.

[69] D. E. Knuth, J. H. Morris, Jr., and V. R. Pratt. Fast pattern matching in strings. *SIAM J. Comput.*, 6(2):323–350, 1977.

[70] R. Kolpakov and G. Kucherov. Searching for gapped palindromes. *Theoret. Comput. Sci.*, 410(51):5365–5373, 2009.

[71] M. Levy and E. Yaakobi. Mutually uncorrelated codes for DNA storage. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 3115–3119, 2017.

[72] M. Lothaire. *Combinatorics on Words*. Addison-Wesley, Reading, Mass., 1983.

[73] R. C. Lyndon and M. P. Schützenberger. The equation $a^M = b^N c^P$ in a free group. *Michigan Math. J.*, 9(4):289–298, 1962.

[74] K. Mahalingam, A. Maity, P. Pandoh, and R. Raghavan. Block reversal on finite words. *Theoret. Comput. Sci.*, 894:135–151, 2021.

[75] J. Massey. Optimum frame synchronization. *IEEE Trans. Commun.*, 20(2):115–119, 1972.

[76] H. Morita, A. J. van Wijngaarden, and A. J. Han Vinck. Prefix synchronized codes capable of correcting single insertion/deletion errors. In *Proceedings of IEEE International Symposium on Information Theory*, page 409. IEEE, 1997.

[77] H. Morita, A. J. van Wijngaarden, and A.J. Han Vinck. On the construction of maximal prefix-synchronized codes. *IEEE Trans. Inform. Theory*, 42(6):2158–2166, 1996.

[78] H. Mousavi. Automatic theorem proving in `Walnut`. Arxiv preprint, available at https://arxiv.org/abs/1603.06017. Software available at https://github.com/hamousavi/Walnut, 2016.

[79] J. Nicholson and N. Rampersad. Improved estimates for the number of privileged words. *J. Integer Sequences*, 21, 2018. Article 18.3.8.

[80] P. T. Nielsen. A note on bifix-free sequences. *IEEE Trans. Inform. Theory*, IT-19:704–706, 1973.

[81] J. Peltomäki. Introducing privileged words: Privileged complexity of Sturmian words. *Theoret. Comput. Sci.*, 500:57–67, 2013.

[82] N. Rampersad, J. Shallit, and M.-w. Wang. Inverse star, borders, and palstars. *Inform. Process. Lett.*, 111:420–422, 2011.

[83] O. Ravsky. On the palindromic decomposition of binary words. *J. Autom. Lang. Comb.*, 8(1):75–83, 2003.

[84] M. Régnier. Enumeration of bordered words, le langage de la vache-qui-rit. *RAIRO-Theor. Inf. Appl.*, 26(4):303–317, 1992.

[85] L. B. Richmond and J. O. Shallit. Counting the palstars. *Electronic J. Combinatorics*, 21(3), 2014. P3.25.

[86] J. Rukavicka. Upper bound for the number of closed and privileged words. *Inform. Process. Lett.*, 156:105917, 2020.

[87] J. Rukavicka. Upper bound for the number of privileged words. *Discrete Math.*, 346(1):113164, 2023.

[88] W. Rytter. A correct preprocessing algorithm for Boyer-Moore string-searching. *SIAM J. Comput.*, 9:509–512, 1980.

[89] L. Schaeffer and J. Shallit. Closed, palindromic, rich, privileged, trapezoidal, and balanced words in automatic sequences. *Electronic J. Combinatorics*, 23(1), 2016. P1.25.

[90] R. Scholtz. Frame synchronization techniques. *IEEE Trans. Commun.*, 28(8):1204–1213, 1980.

[91] J. Shallit. *A Second Course in Formal Languages and Automata Theory*. Cambridge University Press, 2008.

[92] J. Shallit. Hamming distance for conjugates. *Discrete Math.*, 309(12):4197–4199, 2009.

[93] J. Shallit. Fifty Years of Fine and Wilf. Talk presented at the Vrije Universiteit, Amsterdam. https://cs.uwaterloo.ca/~shallit/Talks/vu3.pdf, November 2015.

[94] D. M. Silberger. Borders and roots of a word. *Portugal. Math*, 30:191–199, 1971.

[95] N. J. A. Sloane et al. OEIS Foundation Inc. (2022), The On-Line Encyclopedia of Integer Sequences, https://oeis.org.

[96] C. Stefanović and D. Bajić. On the search for a sequence from a predefined set of sequences in random and framed data streams. *IEEE Trans. Commun.*, 60(1):189–197, 2012.

[97] S. M. H. Tabatabaei Yazdi, H. M. Kiah, R. Gabrys, and O. Milenković. Mutually uncorrelated primers for DNA-based data storage. *IEEE Trans. Inform. Theory*, 64(9):6283–6296, 2018.

[98] A. Thue. Über unendliche Zeichenreihen. *Norske vid. Selsk. Skr. Mat. Nat. Kl.*, 7:1–22, 1906. Reprinted in *Selected Mathematical Papers of Axel Thue*, T. Nagell, editor, Universitetsforlaget, Oslo, 1977, pp. 139–158.

[99] A. Thue. Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen. *Norske vid. Selsk. Skr. Mat. Nat. Kl.*, 1:1–67, 1912. Reprinted in *Selected Mathematical Papers of Axel Thue*, T. Nagell, editor, Universitetsforlaget, Oslo, 1977, pp. 413–478.

[100] J. Wei. The structures of bad words. *European J. Combinatorics*, 59:204–214, 2017.

[101] J. Wei, Y. Yang, and X. Zhu. A characterization of non-isometric binary words. *European J. Combinatorics*, 78:121–133, 2019.

[102] J. Ziv and A. Lempel. A universal algorithm for sequential data compression. *IEEE Trans. Inform. Theory*, 23(3):337–343, 1977.

APPENDICES

Appendix A

Walnut code for Chapter 3

```
def isBorderC1 "((k+l>n) => ((Ai (i<n-l) => T[m+l+i] = T[m+l-k+i]) &
                             (Ai (i<k+l-n) => T[m+i] = T[m+n-k+i])))":

def isBorderC2 "(((k+l<=n) & (l>=k)) => (Ai (i<k) => T[m+l+i] = T[m+l-k+i]))":

def isBorderC3 "(((k+l<=n) & (l<k)) => ((Ai (i<k-l) => T[m+n-k+l+i] = T[m+l+i]) &
                                        (Ai (i<l) => T[m+i] = T[m+k+i])))":

def isBorder "$isBorderC1(k,l,m,n) & $isBorderC2(k,l,m,n) & $isBorderC3(k,l,m,n)":

def isBordered "Ei (2*i<=n & i>=1 & $isBorder(i,l,m,n))":

def isAlternatingE "(Ai ((i!=l & i!= e & i<n-1) =>
                         ($isBordered(i,m,n) <=> ~$isBordered(i+1,m,n))) &
                        (((i!=l) & (i!=e) & (i=n-1)) => ($isBordered(n-1,m,n) <=>
                                                        ~$isBordered(0,m,n))))":

def isAlternatingO "(Ai ((i!=l & i<n-1) =>
                         ($isBordered(i,m,n) <=> ~$isBordered(i+1,m,n))) &
                        (((i!=l) & (i=n-1)) => ($isBordered(n-1,m,n) <=>
                                               ~$isBordered(0,m,n))))":

def hasMNUCE "(Ei,j ((i<j) & (i<n-1 & $isBordered(i,m,n) & $isBordered(i+1,m,n)) &
                    ((j=n-1 & $isBordered(n-1,m,n) & $isBordered(0,m,n)) |
```

```
                        ((j<n-1) & $isBordered(j,m,n) & $isBordered(j+1,m,n))) &
                       $isAlternating(i,j,m,n))) | $isAlternatingE(n,n,m,n)":

def hasMNUCO "Ei (((i<n-1 & $isBordered(i,m,n) & $isBordered(i+1,m,n)) |
                   (i=n-1 & $isBordered(n-1,m,n) & $isBordered(0,m,n))) &
                  $isAlternatingO(i,m,n))":

eval verifyEven "An ((n>=2) => (Ei (i<=2*n) & $hasMNUCE(i,2*n)))":

eval verifyOdd "An ((n>=2) => (Ei (i<=2*n+1) & $hasMNUCO(i,2*n+1)))":
```